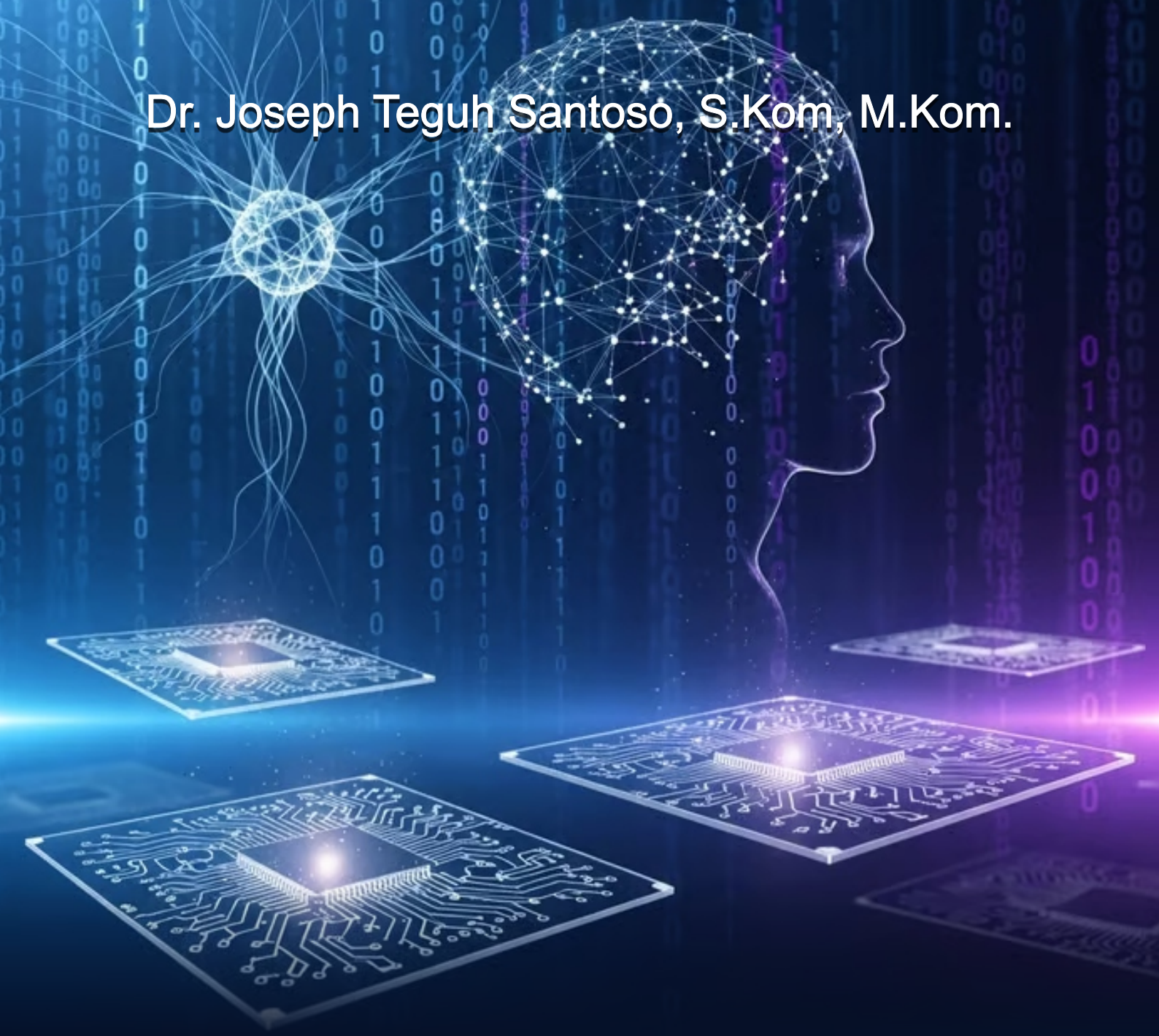
YAYASAN PRIMA AGUS TEKNIK

ETIKA PADA AI (Artificial Intelligence) dan TI (Teknologi Informasi)

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

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



ETIKA PADA AI (Artificial Intelligence) dan TI (Teknologi Informasi)

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 978-634-7227-70-6 (PDF)
9 786347 227706

Etika pada AI (Artificial Intelligence) dan TI (Teknologi Informasi)

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

ISBN : 978-634-7227-70-6 (PDF)

Editor :
Dr. Agus Wibowo, M.Kom, M.Si, MM.

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

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

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

Anggota IKAPI No: 279 / ALB / JTE / 2023

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

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

KATA PENGANTAR

Perkembangan pesat kecerdasan buatan (AI) dan teknologi informasi (TI) telah membawa perubahan fundamental dalam berbagai aspek kehidupan manusia—dari cara kita bekerja, berkomunikasi, hingga mengambil keputusan penting. Namun, di balik kemajuan teknologi yang menjanjikan ini, muncul berbagai tantangan etis yang mesti dipahami dan ditangani dengan bijak. Buku Etika AI dan TI hadir sebagai upaya untuk menjembatani pemahaman mendalam mengenai aspek etika dalam pengembangan dan penerapan AI serta TI.

Buku ini disusun secara sistematis dan komprehensif, dimulai dengan pengenalan konsep etika AI, dilanjutkan dengan pembahasan pendekatan sosioteknis yang mengintegrasikan aspek sosial dan teknis dalam teknologi AI. Pembaca juga akan diajak menyelami isu kritis seperti bias, keberlanjutan, akuntabilitas, dan transparansi dalam penggunaan AI, serta regulasi dan tata kelola yang diperlukan untuk memastikan AI yang bertanggung jawab.

Selain itu, buku ini mengulas teknologi mutakhir seperti transformator, AI generatif, natural language processing (NLP), dan Large Language Models (LLM), serta tantangan etika yang terkait dengan perkembangan tersebut. Ada juga pembahasan mendalam tentang multimodalitas dan risiko yang muncul dari model AI percakapan yang semakin maju, yang membuka perspektif baru dalam kerangka etika AI.

Buku Etika AI dan TI ini dirancang untuk memberikan pemahaman komprehensif mengenai dimensi etis yang melekat pada perkembangan kecerdasan buatan (AI) dan teknologi informasi (TI), dengan fokus pada konteks akademis dan penelitian. Bab pertama menguraikan landasan konseptual etika AI, mengkaji relevansi dan tantangan normative yang muncul seiring dengan pesatnya inovasi teknologi. Selain itu, bab ini juga membahas kritik-kritik filosofis dan praktis yang melandasi perlunya integrasi nilai etis dalam desain dan pengembangan AI.

Beranjak ke bab kedua, pembahasan diarahkan pada pendekatan sosioteknis yang memandang AI tidak sekadar sebagai entitas teknis, melainkan sebagai fenomena yang berinteraksi kompleks dengan struktur sosial. Bab ini menggali elemen-elemen kunci dalam pendekatan tersebut dan menyajikan contoh aplikatif yang relevan bagi kajian interdisipliner. Bab ketiga selanjutnya membahas isu keberlanjutan teknologi AI dan problematika krisis replikasi—suatu fenomena yang berimplikasi pada validitas ilmiah dan keandalan teknologi yang dikembangkan.

Bab keempat secara khusus meneliti problem bias dalam AI, dengan analisis kritis terhadap sumber-sumber bias dan dampaknya terhadap keadilan algoritmik. Bab kelima melanjutkan diskursus ini dengan mengkaji tiga aspek pokok dalam etika AI: keadilan, akuntabilitas, dan transparansi. Pembahasan difokuskan pada dimensi teoritis dan praktis dari ketiga konsep tersebut, yang menjadi kerangka kerja penting dalam pengaturan dan evaluasi sistem AI.

Regulasi dan tata kelola AI menjadi tema utama bab keenam, dimana diuraikan berbagai inisiatif audit dan kebijakan yang menjawab kebutuhan pengawasan teknologi yang semakin kompleks. Bab ketujuh memperkenalkan konsep Explainable AI (XAI), menyoroti metode-metode yang memungkinkan interpretabilitas model AI agar dapat dipertanggungjawabkan secara ilmiah dan etis.

Selanjutnya, bab kedelapan memaparkan struktur dan fungsi arsitektur transformer sebagai fondasi teknologi AI generatif mutakhir, disertai analisis peran umpan balik manusia dalam peningkatan performa model. Bab kesembilan mengkaji bias representasional dalam Natural Language Processing (NLP), menyoroti stereotip dan keterbatasan strategi mitigasi bias, yang penting untuk pengembangan model bahasa yang lebih adil dan akurat. Bab kesepuluh membahas manfaat dan risiko Large Language Models (LLM), termasuk fenomena halusinasi dan strategi mitigasinya, serta refleksi kritis terhadap implikasi sosialnya.

Dalam bab kesebelas, perhatian difokuskan pada transformator visual dan multimodalitas, termasuk teknologi jaringan adversarial generatif dan model difusi, serta integrasi multimoda dalam sistem percakapan AI. Bahasan ini meliputi juga risiko-risiko yang muncul dari kemajuan teknologi tersebut, memberikan wawasan mendalam bagi kajian etika interdisipliner. Sebagai penutup, bab kedua belas menyajikan perspektif futuristik dan tantangan etis yang harus diantisipasi untuk gelombang AI berikutnya, dengan penekanan pada eksposur etika dan pengembangan kebijakan yang progresif dan bertanggung jawab.

Semoga buku ini dapat menjadi sumber pengetahuan dan inspirasi dalam membangun ekosistem AI dan TI yang tidak hanya canggih secara teknologi, tetapi juga berlandaskan nilai-nilai etika yang kuat demi masa depan yang lebih baik.

Semangat dan Selamat membaca…!!!

Semarang, Desember 2025
Penulis

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

"Keberlanjutan teknologi terletak pada bagaimana kita menjaga keseimbangan antara kemampuan dan nilai-nilai manusia. Etika bukan hambatan, melainkan pondasi yang memperkuat inovasi teknologi."

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

DAFTAR ISI

BAB 1
APA ITU ETIKA AI?

1.1 PENDAHULUAN

Dalam bab pengantar ini, kami menyoroti aspek-aspek kunci etika terapan dalam konteks Kecerdasan Buatan (AI). Kami menjelaskan apa yang dimaksud dengan etika AI, tujuannya, dan berbagai cara peneliti dapat berkontribusi untuk memajukan bidang ini. Lebih spesifik lagi, kami ingin menunjukkan perspektif kami sebagai kelompok interdisipliner yang berinteraksi dengan etika AI dari berbagai sudut pandang.

Dengan menyadari kebutuhan ini, kami ingin menawarkan pandangan untuk memahami dan mempelajari Etika AI yang berasal dari kebutuhan, pengalaman, dan keterbatasan kami sendiri sebagai peneliti. Oleh karena itu, dalam buku ini, kami akan menggabungkan keahlian kami dalam etika terapan, Interaksi Manusia-Komputer (HCI), dan Pemrosesan Bahasa Alami (NLP) untuk menyajikan berbagai dimensi isu etika dalam AI.

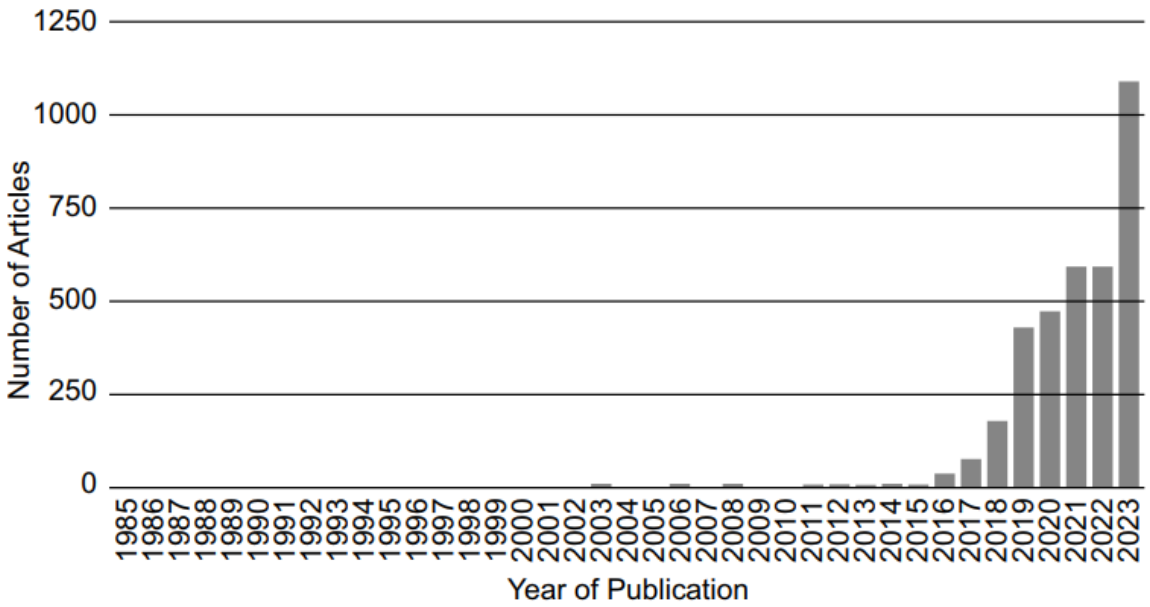
Pertama, kami akan menjelaskan mengapa Etika AI menjadi semakin relevan dan mengapa hal ini merupakan upaya yang menantang. Kemudian, kami akan menjelaskan bagaimana kami memahami etika AI, posisi para kritikus seputar etika AI, dan alur penelitian yang kami ikuti untuk mencari pendekatan dalam mengatasi tantangan etika yang muncul dalam proyek-proyek AI. Sebagai penutup, kami akan menyoroti gagasan-gagasan kunci yang membentuk perdebatan terkini tentang etika AI, yang akan menjadi landasan bagi diskusi-diskusi di bab-bab selanjutnya.

1.2 RELEVANSI DAN TANTANGAN ETIKA AI

Selama dekade terakhir, etika AI telah berevolusi dari bidang yang sedang berkembang dalam Etika Terapan menjadi kebutuhan yang meluas bagi para peneliti, perusahaan, pemerintah, dan pengembang. Masyarakat menuntut standar etika untuk memandu revolusi teknologi yang cepat dan ekspansif yang didorong oleh teknologi AI. Hal ini sejalan dengan bagaimana bioetika menjadi titik fokus dalam Etika Terapan ketika dilema seperti kloning, eutanasia, aborsi, dan pengujian genetik muncul. Dilema-dilema ini membutuhkan para ahli yang tidak hanya berpengetahuan dalam teori etika tetapi juga mampu menerapkannya dalam praktik medis, sehingga menjadikan bioetika sebagai sub-bidang yang inheren interdisipliner.

Etika AI muncul untuk mengatasi tantangan pengembangan dan penerapan teknologi AI di masyarakat, termasuk berbagai skenario aplikasi. Untuk menggambarkan peningkatan relevansinya, perhatikan bagaimana jumlah makalah akademis di Google yang memuat kata "AI" dan "Etika" dalam judulnya telah meningkat sejak tahun 1985 (lihat Gambar 1.1). melakukan pencarian untuk menunjukkan bahwa meskipun mengeksplorasi isu-isu etika dalam AI mungkin tampak lumrah saat ini, kenyataannya tidak selalu demikian. Dalam gambar aslinya, mereka menunjukkan jumlah artikel yang muncul di Google Scholar dengan tag ("etika" atau "etis") dan ("AI" atau "kecerdasan buatan"). Kami melakukan pencarian yang diperbarui dengan kriteria yang sama hingga tahun 2023 (Gambar 1.1) dan mengamati bahwa

tren peningkatan ini terus berlanjut. Konsolidasi relevansi etika AI tampaknya terjadi di mana-mana saat ini, setidaknya di dunia akademis.



Gambar 1.1: Pembaruan (hingga 2023) pencarian Google Scholar.

Yang menarik dari tinjauan historis ini dalam wacana ilmiah adalah asal mula "hype" untuk Etika AI, yang dipicu oleh kasus-kasus terkenal yang terjadi pada tahun 2016, seperti "algoritma rasis" COMPAS (Correctional Offender Management Profiling for Alternative Sanctions). Kasus-kasus nyata secara signifikan memengaruhi integrasi etika ke dalam pengembangan AI. AI bertransisi dari bidang khusus menjadi teknologi arus utama. Dengan evolusinya dan munculnya big data serta kemampuan komputasi yang canggih, dampak sosial AI membuat pertimbangan etika dalam pengembangan dan implementasinya menjadi diskusi etika yang tak terelakkan.

Diskusi ini dengan cepat menjadi menantang karena konsep-konsep dengan tradisi panjang dalam filsafat moral, seperti keadilan, telah mengambil makna dan konteks baru dalam aplikasi AI. Mari kita telusuri lebih dalam kasus COMPAS untuk mengilustrasikan hal ini. Algoritma penilaian risiko yang dikembangkan oleh Equivant (sebelumnya Northpointe, Inc.) ini dirancang untuk memprediksi kemungkinan residivisme seseorang. Tujuan prediksi ini adalah untuk membantu hakim dan memberi informasi kepada petugas pembebasan bersyarat dalam memutuskan apakah terdakwa yang menunggu persidangan terlalu berisiko dan, oleh karena itu, tidak dapat diberikan pembebasan dengan jaminan. yaitu, jika skor residivisme yang diprediksi terlalu tinggi, mereka harus dianggap berisiko. Data pelatihan algoritma ini berasal dari berbagai sumber. Sebagian data pelatihan berasal dari penilaian risiko historis yang dilakukan oleh hakim di Amerika Serikat.

Data tersebut juga mencakup tanggapan dari kuesioner yang diisi oleh terdakwa dan evaluasi oleh petugas pemasyarakatan mengenai persepsi mereka terhadap risiko para pelaku. Model AI yang telah dilatih memberikan skor mulai dari 1 hingga 10 kepada terdakwa, dengan

skor yang lebih tinggi menunjukkan risiko residivisme yang lebih besar. Perdebatan etika dipicu oleh sekelompok jurnalis investigasi dari ProPublica, yang melaporkan bahwa pelaku kejahatan berkulit hitam lebih mungkin dilabeli berisiko tinggi dibandingkan pelaku kejahatan berkulit putih, yang menunjukkan adanya bias dalam algoritma terhadap individu berkulit hitam. ProPublica menganalisis skor risiko yang ditetapkan dari tahun 2013 hingga 2014 dan membandingkannya dengan pengulangan pelanggaran individu dalam dua tahun berikutnya. Misalnya, seorang terdakwa berkulit hitam, B.P. dilabeli berisiko tinggi oleh COMPAS (skor 10/10) meskipun hanya memiliki satu pelanggaran sebelumnya berupa perlawanan terhadap penangkapan tanpa kekerasan dan ia tidak melakukan pelanggaran berikutnya dalam dua tahun.

Sebaliknya, seorang terdakwa berkulit putih, V.P., dilabeli berisiko rendah (skor 3/10) meskipun memiliki riwayat kriminal dua perampokan bersenjata dan satu percobaan perampokan bersenjata dan ia melakukan satu pencurian besar dalam dua tahun. Kasus-kasus ini bukanlah contoh yang terisolasi. Analisis positif palsu dan negatif palsu pada sampel lebih dari 6.000 orang menunjukkan bahwa meskipun algoritma mampu memprediksi residivisme dengan tepat sebanyak 61%, algoritma tersebut menghasilkan kesalahan yang berbeda antar ras. Lihat matriks kebingungan untuk terdakwa Kulit Hitam dan Kulit Putih pada Tabel 1.1 dan tingkat kesalahan berdasarkan ras pada Tabel 1.2. Lebih spesifik lagi, tingkat kesalahan klasifikasi positif palsu—di mana individu diprediksi memiliki risiko residivisme yang lebih tinggi, tetapi mereka tidak mengulangi tindak pidana dalam dua tahun berikutnya—hampir dua kali lebih tinggi untuk terdakwa Kulit Hitam dibandingkan dengan terdakwa Kulit Putih (44,8% vs 23,5%).

Tabel 1.1: Matriks kebingungan untuk terdakwa kulit hitam dan kulit putih

	Terdakwa Black			**Terdakwa White**	
	Risiko Prediksi			**Risiko Prediksi**	
	Rendah	**Tinggi**	**Total**	**Rendah**	**Tinggi**
Tidak mengulangi kejahatan	990	805	1795	1139	349
Mengulangi kejahatan	532	1369	1901	461	505
Total	**1522**	**2174**	**3696**	**1600**	**854**

Sebaliknya, tingkat kesalahan klasifikasi negatif palsu, di mana orang-orang diklasifikasikan dengan risiko residivisme yang lebih rendah, tetapi pada kenyataannya mereka melakukan tindak pidana berulang. jauh lebih tinggi untuk residivisme kulit putih dibandingkan residivisme kulit hitam (47,7% vs 28%). Dari temuan ini, dapat disimpulkan bahwa algoritma tersebut melebih-lebihkan risiko residivisme untuk terdakwa kulit hitam dan meremehkannya untuk terdakwa kulit putih, yang mengarah pada kesimpulan bahwa algoritma tersebut bias terhadap terdakwa kulit hitam.

Menanggapi klaim ProPublica bahwa algoritma tersebut "rasis" dan tidak adil, Equivant berpendapat bahwa penggunaan metrik kesalahan model oleh ProPublica, seperti rasio positif palsu dan negatif palsu, untuk mengevaluasi bias rasial adalah keliru. Equivant berpendapat bahwa pengukuran lain, seperti nilai prediktif positif dan negatif (beserta pelengkapnya), seharusnya digunakan. Argumen utama mereka adalah bahwa rasio positif palsu dan negatif palsu dipengaruhi oleh perbedaan rasio dasar perilaku yang diteliti. Mereka menunjukkan bahwa perbedaan rasio residivisme antara terdakwa kulit hitam dan kulit putih (masing-masing 0,51 dan 0,39) memengaruhi metrik kesalahan, yang mengakibatkan rasio positif palsu yang lebih tinggi di antara terdakwa kulit hitam karena rasio dasar mereka yang lebih tinggi.

Equivant menegaskan bahwa nilai prediktif positif dan negatif, yang menilai akurasi algoritma dalam membuat prediksi positif atau negatif lintas ras, memberikan pengukuran bias rasial algoritma yang lebih akurat. Hal ini mengalihkan fokus dari kesalahan yang berkaitan dengan perilaku aktual ke kesalahan yang berkaitan dengan prediksi spesifik (lihat perhitungan pada Tabel 1.2 dan 1.3). Equivant menyarankan bahwa agar suatu algoritma dianggap adil, algoritma tersebut harus menunjukkan paritas prediktif. Ini berarti algoritma tersebut harus memiliki kemampuan yang sama untuk mengidentifikasi residivis dan non-residivis di antara populasi Kulit Hitam dan Kulit Putih.

Dengan menggunakan nilai prediktif positif (37% untuk terdakwa Kulit Hitam, 40,9% untuk terdakwa Kulit Putih), Equivant menyimpulkan bahwa kemungkinan residivisme di antara terdakwa berisiko tinggi serupa di semua ras. Hal ini, menurut mereka, menunjukkan bahwa algoritma mereka mencapai paritas prediktif. Oleh karena itu, menurut Equivant, klaim ProPublica tidak berdasar karena pola positif palsu (persentase non-residivis yang salah diklasifikasikan sebagai residivis) yang memengaruhi orang kulit hitam dibandingkan orang kulit putih "tidak menunjukkan bukti bias, melainkan merupakan konsekuensi alami dari penggunaan aturan penilaian yang tidak bias untuk kelompok yang kebetulan memiliki distribusi skor yang berbeda".

Kasus hukum, 'State of Wisconsin v. Loomis' 881 N.W.2d 749 di Amerika Serikat, menambahkan preseden hukum pada kasus COMPAS. Terdakwa mengklaim ada pelanggaran hak proses hukumnya dengan menggunakan algoritma penilaian risiko ini. Loomis dan pengacaranya meminta panggilan pengadilan karena mereka tidak dapat memperoleh rincian spesifik tentang bagaimana algoritma tersebut menghasilkan skor individualnya. Menurut pengadilan, COMPAS tidak diharuskan untuk menyajikan rincian tentang algoritmanya, karena mereka diklasifikasikan sebagai rahasia dagang.

Tabel 1.2: Angka-angka ProPublica untuk mendukung klaimnya tentang bias rasial

	Terdakwa Black	**Terdakwa White**
Tingkat Positif Palsu	$805/1795 \times 100(44.8\%)$	$349/1488 \times 100(23.5\%)$
Tingkat Negatif Palsu	$532/1901 \times 100(28\%)$	$461/966 \times 100(47.7\%)$

Tabel 1.3: Angka-angka Equivant untuk mendukung klaimnya tentang paritas prediktif

Kulit Hitam	Kulit Putih
1-PPV (1-1369/2174) * 100 (37%)	(1-505/854) * 100 (40.9%)
1-NPV (1-990/1522) * 100 (35%)	(1-1139/1600) * 100 (28.8%)
PPV: Nilai prediksi positif	
NPV: Nilai prediksi negatif	

Ketika Loomis mengajukan banding atas penggunaan COMPAS, pengadilan mengklaim bahwa adalah akurat untuk mengatakan bahwa hukuman Loomis akan sama persis terlepas dari penggunaan COMPAS. Pengadilan menegaskan kembali bahwa penggunaan COMPAS membantu menguatkan hukuman dan, oleh karena itu, tidak ada campur tangan dalam pengambilan keputusan.

Dalam kasus hukum ini, Mahkamah Agung menolak klaim Loomis, dengan alasan ketersediaan laporan COMPAS bagi Negara dan terdakwa (dengan hasil yang diberikan sebelum persidangan) dan kesempatan bagi terdakwa untuk merevisi tanggapan kuesioner. Evaluasi pengadilan terhadap algoritma tersebut berfokus pada keakuratan tanggapan terdakwa terhadap kuesioner (sebagian dari data masukan algoritma) dan kesempatan untuk memverifikasinya. Selain itu, pengadilan mencatat bahwa kalimat akhir akan tetap tidak berubah terlepas dari prediksi COMPAS.

Para peneliti berpendapat bahwa menilai kewajaran suatu algoritma hanya berdasarkan keakuratan data masukan tidaklah memadai. Pendekatan ini mengabaikan aspek-aspek penting lainnya dari pemrosesan dan analisis data. Washington berpendapat bahwa berfokus hanya pada satu aspek kualitas data mengabaikan kompleksitas yang lebih luas yang terlibat dalam mengevaluasi algoritma, terutama dalam sektor publik. Misalnya, kekhawatiran tentang visibilitas keputusan dan kesulitan dalam meneliti prosedur terlindungi dalam ketidakjelasan seputar algoritma. Oleh karena itu, sekilas, apa yang tampak sebagai isu keadilan atau kesetaraan juga berkaitan dengan transparansi.

Perhatian krusial lainnya terkait COMPAS terkait dengan jenis pertanyaan yang terdapat dalam kuesioner penilaian risiko, yang berfungsi sebagai masukan bagi algoritma. Beberapa pertanyaan ini berpotensi menimbulkan isu terkait representasi yang adil dan stigmatisasi, misalnya:

- Berdasarkan pengamatan pemeriksa, apakah orang ini diduga atau diakui sebagai anggota geng?
- Berapa banyak teman/kenalan Anda yang pernah ditangkap?
- Apakah Anda pernah diskors atau dikeluarkan dari sekolah?

Mempertimbangkan pertanyaan-pertanyaan ini sebagai masukan yang relevan mungkin secara inheren mencerminkan dan melestarikan bias, yang mempertanyakan, dari perspektif lain, kewajaran penerapan algoritma.

Oleh karena itu, kita menghadapi skenario yang menantang untuk menetapkan kriteria kewajaran. Di satu sisi, Equivant membenarkan hasil tersebut sebagai tidak rasis berdasarkan paritas prediktif dan analisis teknis, dengan mengklaim bahwa algoritma tersebut sama adilnya

dalam memprediksi lintas ras. Di sisi lain, argumen ProPublica berpusat pada hasil, menunjukkan titik buta etika COMPAS: COMPAS secara tidak proporsional memengaruhi pelaku kejahatan berkulit hitam dengan melebih-lebihkan risiko mereka untuk mengulangi kejahatan, meskipun ras tidak secara eksplisit dimasukkan sebagai variabel dalam prediksi. Mahkamah Agung kemudian memutuskan untuk membatasi kriteria tersebut pada pengukuran data masukan yang akurat, sementara beberapa data masukan dapat dipertanyakan karena legitimasinya.

COMPAS menyoroti kompleksitas perdebatan tentang kewajaran dalam AI dengan mempertimbangkan berbagai dimensi seperti aspek teknis, etika, dan hukum. Hal ini menunjukkan tantangan dalam memahami keadilan melalui satu sudut pandang tunggal karena berkaitan dengan banyak aspek, seperti akurasi, bias, aspek prosedural, diskriminasi, dan transparansi. COMPAS menggambarkan bagaimana diskusi tentang keadilan algoritmik dapat dengan cepat berkembang menjadi perdebatan tanpa akhir tentang makna keadilan.

Tantangan dalam membahas keadilan dalam AI juga melibatkan kesulitan dalam mengoperasionalkan dan mengimplementasikan gagasan teknis tentang keadilan pada model algoritmik. Masukan etis seringkali mencerminkan ketimpangan struktural dalam masyarakat, kriteria perundang-undangan yang ada dapat mengabaikan atau memprioritaskan definisi spesifik untuk mengevaluasi keadilan, dan kriteria yang diperlukan untuk mengevaluasi keberhasilan pengembangan AI dapat memengaruhi ukuran keadilan secara berbeda. Lebih lanjut, berbagai dimensi keadilan ini seringkali dirancang dan dibatasi pada domain disiplin ilmunya. Meskipun secara teoritis dapat memengaruhi dimensi lain, dimensi-dimensi tersebut jarang dikonseptualisasikan sebagai satu kesatuan yang terintegrasi mengingat keterbatasan praktis yang ada.

Analisis kasus ini menyoroti tidak hanya tugas kompleks dalam menjawab pertanyaan etika terkait AI, tetapi juga semakin signifikannya etika AI dalam beberapa tahun terakhir. Hal ini terutama berlaku untuk topik-topik yang berdampak langsung pada masyarakat. Aplikasi AI digunakan untuk meningkatkan efisiensi pertanian, mempelajari kebakaran hutan dan gempa bumi, serta memantau pasang surut air laut. AI juga meluas ke bidang-bidang yang memiliki hubungan lebih langsung dengan kebutuhan dan kesejahteraan manusia, seperti merawat pasien demensia, mengelola dan mengawasi karyawan, melacak kemajuan akademik mahasiswa, memantau individu yang terinfeksi selama pandemi COVID-19, untuk peta navigasi, chatbot asisten, dan sistem peradilan pidana, seperti kasus COMPAS.

Tidak diragukan lagi bahwa etika AI kini sedang digembar-gemborkan; namun, terdapat pertanyaan-pertanyaan etika mendasar yang berakar pada kekhawatiran dari perdebatan etika yang telah berlangsung lama, seperti isu-isu tentang keadilan, transparansi, dan tanggung jawab. Oleh karena itu, salah satu tantangan inti adalah menerjemahkan perdebatan etika ini ke dalam praktik yang bermakna. Mengingat teknologi AI telah menjadi bagian integral dari berbagai aspek kehidupan sehari-hari, mulai dari interaksi pribadi dengan asisten virtual seperti Siri dan Alexa hingga sistem yang lebih kompleks seperti kendaraan otonom dan kepolisian prediktif, adopsi yang meluas ini memerlukan pemeriksaan menyeluruh terhadap implikasi etika teknologi AI terhadap masyarakat. Hal ini juga memerlukan pertimbangan

cermat tentang bagaimana para profesional yang mengembangkan AI mengintegrasikan dan memahami etika AI.

Seiring berkembangnya teknologi dan menjadi bagian integral dari masyarakat, kesadaran akan pentingnya etika dalam pengembangan AI terus tumbuh dalam memengaruhi pengambilan keputusan, penemuan ilmiah, dan pemecahan masalah. Meskipun penelitian etika AI meningkat pesat, menerjemahkan prinsip-prinsip etika ke dalam praktik operasional yang bermakna tetap menjadi tantangan yang signifikan. Misalnya, banyak artikel penelitian, laporan, buku, dan pedoman mendefinisikan prinsip-prinsip untuk AI yang etis, seperti keadilan. Bagaimana mereka mendefinisikan keadilan mungkin memiliki kesamaan tertentu, seperti menghindari bias atau mencegah diskriminasi. Namun, prinsip-prinsip yang menganjurkan "keadilan" ini sering kali tidak memiliki panduan khusus untuk mengatasi masalah terkait keadilan.

Dalam kasus COMPAS yang dibahas sebelumnya, prinsip yang menganjurkan "menghindari bias" tidak dapat diterjemahkan ke dalam praktik-praktik spesifik ketika tiga domain keadilan, teknis, etika, dan hukum, sedang dibahas, terjalin, dan berinteraksi satu sama lain. Oleh karena itu, sebagian tantangan yang kita hadapi saat ini tentang etika AI tidak lagi berakar pada pembuktian relevansinya, melainkan pada penjebolan kesenjangan praktis, metodologis, dan profesional agar etika AI menjadi bernilai nyata, yaitu menawarkan pendekatan, kerangka kerja, dan ukuran konkret yang menjadi bagian integral yang kohesif dari pengembangan dan implementasi AI.

Namun sebelum melanjutkan pembahasan kita, penting untuk memahami apa yang dimaksud dengan etika AI agar dapat meyakinkan mengapa AI harus dipelajari.

1.3 APA YANG AKAN KITA PAHAMI SEBAGAI ETIKA AI?

Untuk mendefinisikan etika AI, pertama-tama kita perlu mengetahui apa itu etika. Filsafat moral adalah bidang yang mempelajari etika, yang berkaitan dengan fenomena moral, yaitu pertanyaan normatif tentang apa yang seharusnya dilakukan orang, berdasarkan alasan dan penilaian yang memungkinkan seseorang untuk berargumen kapan suatu tindakan dapat dianggap benar atau salah secara moral. Dalam studi sistematis fenomena moral ini, terdapat tiga bidang studi utama: metaetika, etika normatif, dan etika terapan.

Metaetika membahas pertanyaan mendasar tentang masalah etika, dengan kata lain, hakikat moralitas. Metaetika menawarkan definisi untuk konsep-konsep dasar seperti kebaikan atau keburukan, objektivitas moralitas, atau makna dan asal-usul penilaian moral. Dengan demikian, metaetika mempelajari bagaimana kita berpikir, mengetahui, dan mengonseptualisasikan moralitas, yang menginformasikan etika normatif.

Etika normatif mengeksplorasi prinsip-prinsip universal untuk memandu tindakan dari berbagai perspektif teoretis seperti konsekuensialisme, etika kebajikan, deontologi, dan etika kepedulian. Bidang ini mengkaji moralitas tindakan, motivasi, dan sifat karakter dengan menjawab pertanyaan seperti "Apa yang membuat pembunuhan salah secara moral?" atau "Apakah mencuri pernah dapat diterima secara moral?" Bidang ini bertujuan untuk mengidentifikasi prinsip-prinsip umum yang membenarkan tindakan moral.

Etika terapan membahas pertanyaan moral praktis terkait isu-isu spesifik seperti hak-hak hewan, perubahan iklim, dan pengembangan AI. Tidak seperti etika normatif yang berfokus pada prinsip-prinsip universal, etika terapan mempertimbangkan implikasi praktis moralitas dan dampak sosialnya. Cabang filsafat ini memperluas diskusi ke situasi dan konteks tertentu, termasuk ranah etika AI.

Etika AI mencakup beberapa aspek pengembangan AI. Misalnya, diskusi sering kali berfokus pada roboetika, yang berkaitan dengan perangkat lunak AI yang terintegrasi ke dalam mesin fisik yang berinteraksi dengan lingkungannya, terutama melalui sensor, yang mengarah pada diskusi tentang hak-hak robot. Demikian pula, terdapat bidang khusus untuk interaksi manusia-robot, yang mengkaji hubungan sosial dan psikologis kita dengan robot serta perdebatan aplikasi spesifik untuk robot sosial, seperti robot perawatan dan robot seks atau aplikasi lain di militer dan sektor kesehatan. Diskusi tentang etika mesin juga erat kaitannya, yang berfokus pada mesin bermoral dan agen moral buatan.

Area perdebatan lain dalam etika AI berkisar pada kemajuan spesifik di bidang AI dan implikasi sosialnya. Misalnya, penelitian etika tentang kendaraan otonom sering kali merujuk pada dilema etika yang telah lama dikenal sebagai "Masalah Troli". Dilema ini melibatkan keputusan apakah akan menarik tuas untuk mengalihkan troli dari jalur yang akan membunuh lima orang ke jalur lain yang hanya akan membunuh satu orang. Hipotesis ini sering digunakan sebagai jenis eksperimen pikiran untuk memperdebatkan pilihan yang tepat. Dilema ini bertujuan untuk mengeksplorasi intuisi kita tentang bertindak (membunuh satu orang) atau membiarkan sesuatu terjadi (membiarkan lima orang terbunuh).

Skenario hipotetis ini telah dimasukkan dalam diskusi tentang desain etis kendaraan otonom, dengan mempertimbangkan berbagai faktor dan memprioritaskan perbedaan budaya saat memprogram dan melatih kendaraan ini. Jadi, ketika kendaraan otonom ini mengalami kecelakaan, misalnya, sebagian dari penelitian ini mengkaji bagaimana membuat mereka "menabrak secara moral".

Bidang fokus terkemuka lainnya mencakup AI dan layanan kesehatan, yang diskusinya berkisar seputar algoritma klasifikasi untuk mendiagnosis pasien kanker, penggunaan AI untuk diagnosis kesehatan mental, dan isu-isu etika terkait penggunaan biomarker digital dan AI untuk mendeteksi demensia dini. AI juga diterapkan dalam pendidikan, misalnya melalui aplikasi seluler berbasis AI. Dampak pedagogis umum AI juga sedang dipelajari. Selain itu, otomatisasi menimbulkan risiko terhadap ketenagakerjaan, yang menyebabkan polarisasi ketenagakerjaan. Hal ini ditandai dengan pertumbuhan upah yang lambat bagi pekerja berketerampilan rendah, yang semakin tergantikan oleh produktivitas berbasis AI, sementara pekerja berpendidikan tinggi mengalami peningkatan upah.

Dengan demikian, di sini kita akan memahami Etika AI sebagai sub-bidang Etika Terapan yang berkaitan dengan pembentukan praktik yang baik untuk pengembangan dan implementasi AI yang etis. Sebagai bidang multi- dan interdisipliner, Etika AI mengambil wawasan dari berbagai disiplin ilmu, termasuk namun tidak terbatas pada etika, sosiologi, ilmu komputer, teknik, filsafat, psikologi, kedokteran, dan hukum. Mempelajari etika AI membutuhkan kolaborasi antar profesional dari berbagai bidang untuk mengembangkan

panduan dan metodologi yang mendorong hubungan etika yang harmonis antara AI dan masyarakat.

1.4 ADAKAH SUARA-SUARA KRITIS TERKAIT ETIKA AI?

Seperti semua disiplin ilmu, etika AI menghadapi tantangan dan keterbatasan. Salah satu kritik utama adalah efektivitas pendefinisian prinsip-prinsip AI, sebuah strategi yang umum diadopsi oleh organisasi, negara, dan lembaga multilateral di seluruh dunia. Beberapa penulis telah mengkritik "ketidakbergunaan" prinsip dan pedoman etika AI. Mittelstadt, misalnya, berpendapat bahwa alih-alih menawarkan rekomendasi yang konkret dan terarah, "banyak inisiatif, terutama yang disponsori oleh industri, telah dikarakterisasikan sebagai sinyal kebajikan belaka yang dimaksudkan untuk menunda regulasi dan secara preemptif memfokuskan debat pada masalah abstrak dan solusi teknis".

Perspektif ini menyatakan bahwa standar etika bersifat abstrak, mengabaikan tantangan normatif dan politis dari konsep-konsep kunci AI seperti privasi dan keadilan. Hagendorff berpendapat serupa, mengklaim bahwa "etika AI, atau etika secara umum, tidak memiliki mekanisme untuk memperkuat klaim normatifnya sendiri". Hagendorff menekankan bahwa berdasarkan penerapan prinsip-prinsip etika, lembaga dapat mengambil jalan pintas dan menetapkan pedoman etika mereka sendiri, yang mendorong ilusi pengaturan diri. Oleh karena itu, penerapan prinsip-prinsip etika AI sebagai kedok kepatuhan etika, yang dikenal sebagai pencucian etika, merupakan suatu kekhawatiran. Bietti menggambarkan fenomena ini dalam konteks AI sebagai situasi di mana 'etika' semakin dikaitkan dengan upaya pengaturan diri perusahaan teknologi dan perilaku etis yang dangkal.

Baru-baru ini, Munn mengkritik etika AI karena memberikan "prinsip-prinsip yang tidak bermakna, prinsip-prinsip yang terisolasi, dan prinsip-prinsip yang tidak bergigi, sebuah celah antara prinsip dan praktik". Ia menekankan bahwa prinsip-prinsip etika AI menawarkan panduan yang ambigu, yang memungkinkan praktik-praktik yang ada untuk terus mempertahankan status quo industri. Munn juga menyoroti keengganan para insinyur untuk terlibat dengan pertanyaan-pertanyaan etika, yang merupakan gejala dari masalah yang lebih besar dan lebih luas dalam industri teknologi, yaitu menjadikan "AI yang tidak etis sebagai produk sampingan logis dari industri yang tidak etis".

Serupa dengan itu, mewawancarai 40 pengembang AI dan menemukan bahwa hampir setengah dari mereka melaporkan bahwa menjadi seorang pengembang "tidak etis maupun tidak etis". Apresiasi netral terhadap profesi mereka ini memandang pengembangan AI bukan sebagai upaya yang etis, melainkan sebagai tindakan yang dilakukan untuk mendapatkan bayaran dan tugas-tugas yang berkaitan dengan efisiensi, optimasi, dan hal-hal yang lebih teknis. keputusan. Masalah utama di sini adalah bahwa persepsi AI sebagai netral secara etika sering kali berasal dari kurangnya pengetahuan dan kesadaran tentang apa yang dapat diidentifikasi sebagai aspek etika dari pengembangan AI.

Lebih lanjut, penelitian kami sendiri telah menemukan bukti bahwa pengembang dan peneliti AI cenderung berpikir bahwa etika memberi tahu kita apa yang tidak boleh dilakukan. Banyak peneliti dan pengembang menyoroti bahwa mereka menganggap etika sebagai hal

yang negatif, sebagai pemaksaan yang membatasi penelitian dan inovasi, menjadikan etika lebih sebagai hambatan daripada sekutu. Pandangan ini umumnya dimunculkan oleh orang-orang ketika hubungan utama mereka dengan etika adalah melalui komite etika atau AI yang berprinsip, yang mendorong pemahaman etika yang terbatas. Meskipun mengakui relevansi etika AI, kekhawatiran utama mereka didasarkan pada kurangnya pedoman dan mekanisme konkret untuk menerapkan etika dalam praktik.

Oleh karena itu, kritik-kritik ini menyoroti pentingnya mengembangkan pendekatan baru yang lebih praktis dan kontekstual serta yang menyediakan struktur yang lebih kuat agar prinsip-prinsip tersebut efektif. Hal ini menantang, tidak hanya karena etika AI merupakan bidang baru, tetapi juga karena kolaborasi interdisipliner yang dibutuhkan untuk menghasilkan alternatif, serta lanskap pengembangan AI yang terus berkembang. Dengan demikian, lebih dari sekadar batasan semata, kritik terhadap prinsip-prinsip etika AI ini menuntut cara yang lebih komprehensif dan holistik untuk memahami tantangan yang ditimbulkannya.

Dengan mengakui kritik-kritik ini, menjadi penting untuk membahas mengapa bidang studi ini penting dan mengapa relevan bagi berbagai profesional dan akademisi untuk memahaminya. Jawaban kami didasarkan pada fakta bahwa etika dapat dipahami tidak hanya sebagai proses pengambilan keputusan individu, tetapi juga sebagai "kebutuhan untuk membenarkan atau menjelaskan diri kita kepada orang lain. Etika adalah studi tentang tindakan apa yang benar-benar dapat dipertahankan di bawah pengawasan."Dalam bidang interdisipliner seperti AI, mengomunikasikan dan membenarkan keputusan tentang pengembangan dan implementasi AI dengan tepat semakin penting karena AI semakin banyak diadopsi ke dalam berbagai domain di masyarakat.

1.5 MENGINTEGRASIKAN ETIKA KE AI

Menanggapi kritik terhadap etika AI ini, para peneliti telah menganjurkan penanaman etika ke dalam AI. Hal ini dapat dicapai dengan memastikan bahwa sistem AI mewujudkan nilai-nilai tertentu, mengikutsertakan ahli etika ke dalam tim pengembangan AI, atau mengintegrasikan etika ke dalam kursus pembelajaran mesin.

Peneliti lain, seperti Johnson dan Verdicchio, mengkritik gagasan sederhana tentang "menanamkan" etika ke dalam AI, yang mempersoalkan maknanya. Mereka mengklaim bahwa sekadar menambahkan etika ke dalam teknologi, yaitu, "Etika + AI = AI Etis," adalah perspektif yang berasumsi bahwa prinsip-prinsip etika dapat langsung dikodekan ke dalam AI, mengubahnya menjadi AI etis, sehingga melakukan kekeliruan aditif.

Argumen penulis adalah bahwa agar penambahan AI dan etika dapat dimungkinkan, kedua area ini harus memiliki karakteristik ontologis yang sama yaitu, keduanya harus memiliki sifat yang kompatibel. Namun, karena AI memiliki basis komputasional, hal ini menyiratkan bahwa etika juga harus dikomputasi agar dapat diintegrasikan dengan AI. Ini berarti prinsip-prinsip etika dapat benar-benar ditangkap dalam bentuk komputasional, yang tampaknya tidak masuk akal. Perbedaan ini menimbulkan keraguan tentang kelayakan penerjemahan langsung prinsip-prinsip etika ke dalam algoritma komputasional.

Oleh karena itu, mereka mengusulkan bahwa pemahaman yang lebih luas tentang AI sebagai bagian dari sistem sosioteknis dapat menghindari kekeliruan aditif ini. Pemahaman ini melibatkan pertimbangan bahwa AI bukan sekadar seperangkat alat komputasional, tetapi beroperasi dalam jaringan hubungan manusia, norma-norma sosial, dan praktik organisasi yang kompleks. Dengan demikian, pertimbangan etis dalam AI melampaui komputasi belaka dan melibatkan konteks sosioteknis yang lebih luas di mana sistem AI berada. Perspektif ini menggeser fokus dari upaya menjadikan AI etis secara intrinsik menjadi pertimbangan bagaimana praktik AI memengaruhi dan dipengaruhi oleh nilai-nilai dan norma-norma sosial.

Dalam hal ini, etika bukan berarti memberi tahu kita apa yang tidak boleh dilakukan, melainkan membantu kita membangun sudut pandang yang menganggap AI sebagai alat sosioteknis yang dapat dikembangkan dan diimplementasikan dengan lebih baik ketika pertimbangan etika melekat pada tuntutan membangun dan menggunakan sistem AI. Oleh karena itu, menciptakan AI melibatkan lebih dari sekadar solusi teknis; hal ini menuntut keterlibatan dengan dimensi etika dari domain dan industri tempat AI diterapkan. Para ahli AI dipanggil untuk tidak hanya mengembangkan teknologi tetapi juga terlibat secara kritis dengan implikasi etika dan sosial yang lebih luas dari pekerjaan mereka, alih-alih menanamkan etika ke dalam ontologi komputasional AI.

Dengan demikian, etika AI krusial tidak hanya untuk memastikan bahwa sistem AI dikembangkan dan diimplementasikan secara bertanggung jawab, tetapi juga untuk menumbuhkan budaya di mana pertimbangan etika dipandang fundamental dan melekat pada proses inovasi teknologi. Perubahan budaya ini menantang para pengembang, pembuat kebijakan, dan pemangku kepentingan untuk terlibat secara kritis dengan implikasi etika dari pekerjaan mereka, mendorong AI yang adil dan bermanfaat secara sosial. Pentingnya Etika AI terletak pada kemampuannya untuk menjembatani kesenjangan antara kapabilitas teknologi dan nilai-nilai sosial, memastikan bahwa AI berfungsi sebagai alat untuk transformasi sosial yang positif, alih-alih sumber pertikaian dan perpecahan. Untuk itu, beragam metodologi sosioteknis interdisipliner (lihat Bab 2, bagian 2.2) dapat digunakan untuk terlibat dalam penciptaan AI yang etis.

1.6 APA SAJA PERHATIAN UTAMA DALAM ETIKA AI?

Untuk membingkai diskusi kita tentang perhatian utama dalam Etika AI, kita akan menggunakan metode yang umum digunakan, meskipun banyak dikritik, untuk menyusun diskusi dan praktik dalam etika AI di sekitar prinsip-prinsip. Kami memilih strategi ini terutama untuk menghubungkan dengan penelitian sebelumnya, yang memungkinkan kami meninjau kembali konsep-konsep penting dan mengidentifikasi perbedaan serta hubungan di antara keduanya.

Berbagai organisasi telah menetapkan prinsip-prinsip yang dimaksudkan untuk memandu pengembangan AI. Pendekatan berbasis prinsip ini telah diadopsi oleh kelompok profesional seperti IEEE, perusahaan seperti Google, IBM, Telefonica, dan Microsoft, serta pemerintah dan lembaga multinasional, termasuk UNESCO, OECD, dan Perserikatan Bangsa-Bangsa. Pada awal tahun 2024, 42 negara telah berkomitmen pada prinsip-prinsip AI OECD,

dan resolusi PBB tentang AI telah diadopsi dengan suara bulat oleh seluruh 193 negara anggota. Beberapa akademisi telah menyusun ringkasan dan analisis prinsip-prinsip ini, sebagaimana dirinci dalam Tabel 1.4.

Banyak artikel tinjauan ini mengidentifikasi tema-tema yang berulang, seperti privasi, transparansi, akuntabilitas, dan keadilan. Namun, tinjauan tertentu menyoroti elemen-elemen tertentu lebih menonjol daripada yang lain, dipengaruhi oleh sumber yang mereka periksa atau sektor yang mereka amati. Laporan-laporan ini mempertimbangkan perspektif dari akademisi, pemerintah, atau industri, menawarkan berbagai sudut pandang dan aspek untuk dipertimbangkan.

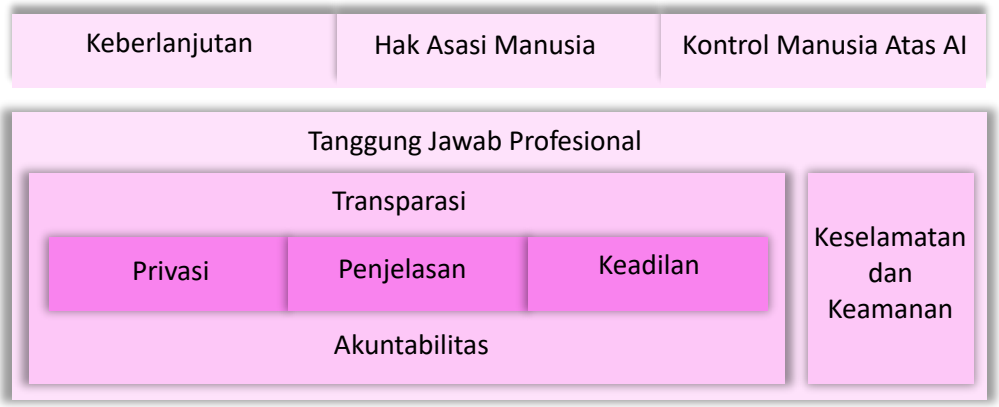
Berdasarkan survei-survei ini, kami mengusulkan kategorisasi untuk menganalisis prinsip-prinsip AI (lihat Gambar 1.2). Kategorisasi ini merupakan hasil dari tinjauan literatur, studi diagnostik kualitatif, dan proses partisipatif yang dilakukan di Pusat AI Nasional, CENIA, di Chili. Kami membagi prinsip-prinsip menjadi dua kategori utama. Di bagian atas, kami menempatkan tiga prinsip inti yang mencakup implikasi etis yang lebih luas dalam AI: keberlanjutan, hak asasi manusia, dan kendali manusia. Ketiga prinsip ini terdiri dari dampak, risiko, dan kekhawatiran yang terkait dengan aspek-aspek fundamental pembangunan manusia berdasarkan penghormatan terhadap perkembangan manusia, otonomi, dan martabat. Oleh karena itu, rangkaian pertama ini merupakan landasan bagi diskusi berbasis prinsip. Di sini, kami menyertakan:

- **Keberlanjutan**, yang mencakup jaminan kondisi kehidupan yang vital dan perlindungan lingkungan untuk generasi mendatang. Prinsip ini membahas dampak lingkungan dari siklus hidup AI, yang mencakup penggunaan energi dan air yang signifikan dalam pengembangan dan penerapan AI, serta ekstraksi sumber daya alam yang ekstensif untuk infrastruktur AI. Inisiatif AI yang etis harus menilai konsumsi energi dan dampak keseluruhannya terhadap lingkungan, berupaya mengurangi jejak karbon, mengoptimalkan efisiensi energi, dan memanfaatkan sumber energi terbarukan untuk infrastrukturnya. Beberapa penulis juga menganggap gagasan komunitas berkelanjutan sebagai aspek fundamental dari dampak keberlanjutan AI, yang melibatkan evaluasi dan peninjauan berkelanjutan terhadap dampak yang diantisipasi dan aktual terhadap komunitas yang terdampak AI, termasuk konsekuensi yang diinginkan dan tidak diinginkan.

Tabel 1.4: Tinjauan dan ringkasan prinsip penelitian Etika AI

Tinjauan	Prinsip
Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., and Srikumar, M. (2020). Principled AI: A Map of Ethical and Rights-Based Approaches to Principles for AI.	Privasi, akuntabilitas, keadilan, keamanan, tanggung jawab profesional, promosi nilai-nilai kemanusiaan, transparansi, kendali manusia atas teknologi.
Khan, A. A., Badshah, S., Liang, P., Khan, B., Waseem, M., Niazi, M., and Akbar, M. A. (2021). Ethics of AI: A	Transparansi, privasi, akuntabilitas, kesetaraan, otonomi, dapat dijelaskan, keadilan, tidak merugikan, martabat

Systematic Literature Review of Principles and Challenges.	manusia, kebaikan, tanggung jawab, keselamatan, keamanan data, keberlanjutan, kebebasan, solidaritas, kemakmuran, efektivitas, keakuratan, prediktabilitas, interpretabilitas.
Jobin, A., Ienca, M., dan Vayena, E. (2019). Lanskap Global Etika AI Pedoman	Transparansi, keadilan dan kewajaran, tidak melakukan hal yang merugikan, akuntabilitas, privasi, berbuat baik, kebebasan dan otonomi, kepercayaan, keberlanjutan, martabat, dan solidaritas.
Zeng, Y., Lu, E., dan Huangfu, C. (2018). Menghubungkan Prinsip Kecerdasan Buatan.	Kemanusiaan, kolaborasi, berbagi (kesetaraan), keadilan, transparansi, privasi, keamanan, perlindungan, akuntabilitas, AGI/ASI.
Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., dan Vayena, E. (2018). AI4People—Kerangka Etis untuk Masyarakat AI yang Baik: Peluang, Risiko, Prinsip dan Rekomendasi.	Kebaikan, tidak merugikan, otonomi, keadilan, dapat dijelaskan
Smit, K., Zoet, M., dan Van Merten, J. (2018). Tinjauan Prinsip-Prinsip AI dalam Praktek.	Peningkatan kualitas manusia, kebaikan hati, keandalan, berpusat pada manusia. Otonomi, kesetaraan (desain dan pelaksanaan), ketertelusuran, martabat manusia, hak asasi manusia, transparansi, demokratisasi, privasi, keamanan, keamanan (desain dan pelaksanaan), kolaborasi, tanggung jawab, pemahaman, penggunaan data yang bertanggung jawab, akurasi, serta pendidikan dan promosi.



Gambar 1.2: Kategorisasi prinsip-prinsip AI.

- **Hak Asasi Manusia** telah digunakan sebagai kerangka kerja untuk menangkap spektrum luas dampak positif dan negatif yang dapat ditimbulkan AI dalam kehidupan manusia. Melindungi hak asasi manusia menjadi penting dalam menanggapi disparitas yang diamati dalam metrik kinerja seperti tingkat kesalahan di antara kelompok gender atau ras yang berbeda, khususnya di bidang yang memengaruhi akses layanan kesehatan, proses hukum, dan keputusan rekrutmen. Resolusi PBB pertama tentang AI menekankan pentingnya menegakkan hak asasi manusia untuk memastikan bahwa sistem AI aman, terlindungi, dan tepercaya. Hak asasi manusia mencakup hak atas kesetaraan, martabat, kebebasan bergerak, kepemilikan properti, dan akses ke layanan kesehatan dan pekerjaan, antara lain. Para akademisi telah menyoroti peran penting hak asasi manusia sebagai kekuatan universal dan normatif, menunjukkan signifikansinya dalam memandu praktik bahkan ketika undang-undang AI nasional yang spesifik tidak ada. Pendekatan terkait lainnya menekankan nilai-nilai kemanusiaan dan perkembangannya, menggarisbawahi peran AI dalam mendukung kemajuan peradaban manusia. Hal ini juga berkaitan dengan konsep berorientasi komunitas yang terkait dengan prinsip keberlanjutan.
- **Kendali manusia atas AI** menekankan perlunya pengawasan manusia terhadap sistem AI sebagai langkah penting untuk mempertahankan otonomi manusia. Hal ini melibatkan pencegahan AI dari pengambilan keputusan kritis secara otonom, menjamin kapasitas untuk melakukan intervensi dalam operasi AI sesuai kebutuhan, dan memastikan bahwa wewenang pengambilan keputusan tetap berada di tangan manusia. Meskipun prinsip ini mungkin tampak berlaku khususnya untuk sistem AI yang sepenuhnya otonom, relevansinya meluas ke situasi di mana manusia terlibat dalam proses yang dimediasi oleh AI. Dalam konteks seperti itu, penting untuk menentukan apakah individu, terutama mereka yang terdampak AI, dapat melakukan intervensi secara efektif. Intervensi tersebut dapat berkisar dari meninjau dan memberikan umpan balik kepada sistem hingga menantang keputusan AI dan memilih keluar dari sistem AI. Selain itu, penting untuk memeriksa apakah para pengambil keputusan yang memanfaatkan informasi yang dihasilkan AI, seperti hakim yang menggunakan skor risiko residivisme, dapat menilai hasil ini secara kritis, mengakui

keterbatasan mereka, dan mempertimbangkan konteksnya. Penelitian telah menunjukkan bahwa orang cenderung terlalu mempercayai hasil yang dihasilkan AI. Bahkan sebelum maraknya pembelajaran mesin, para peneliti telah mendokumentasikan bagaimana otomatisasi terlalu memengaruhi pengambilan keputusan manusia, yang mempersulit implementasi efektif prinsip ini dan implikasinya terhadap akuntabilitas. Lebih lanjut, prinsip ini bertujuan untuk mempertahankan otonomi dengan memastikan bahwa individu dapat membuat keputusan sendiri tanpa dipengaruhi secara berlebihan oleh AI, misalnya melalui persuasi atau dorongan yang didorong oleh AI. Prinsip ini juga menganjurkan persyaratan persetujuan pengguna sebelum berinteraksi dengan sistem AI dan untuk melibatkan pengguna dan pemangku kepentingan lainnya dalam tahap pengembangan dan umpan balik siklus hidup AI.

Selanjutnya, kategorisasi kami menguraikan berbagai prinsip AI di bawah tema utama tanggung jawab profesional. Hal ini menekankan pentingnya individu di balik teknologi AI dan tanggung jawab mereka dalam proses pengambilan keputusan yang memengaruhi implikasi AI terhadap keberlanjutan, hak asasi manusia, dan kendali manusia atas AI, serta bagaimana prinsip-prinsip lain diimplementasikan di seluruh siklus hidup AI. Dalam konteks tanggung jawab profesional, kami membedakan antara masalah keselamatan dan keamanan dengan prinsip-prinsip lain yang lebih terpengaruh oleh kurangnya transparansi metodologis dan model dalam pengembangan AI. Karena alasan ini, kami menonjolkan transparansi sebagai prinsip utama yang mencakup dan saling terkait dengan semua prinsip yang kami tempatkan di dalamnya, seperti akuntabilitas, privasi, keterjelasan, dan keadilan. Deskripsi terperinci dari masing-masing prinsip ini adalah sebagai berikut:

- **Tanggung jawab profesional** mengacu pada peran penting individu yang terlibat dalam pengambilan keputusan, perancangan, pengembangan, dan penerapan AI. Hal ini menekankan harapan bagi para profesional ini untuk mengatasi berbagai masalah, mulai dari memastikan kinerja sistem dengan mempertimbangkan dampak jangka panjang dari keputusan mereka hingga prinsip-prinsip inti seperti keberlanjutan dan hak asasi manusia. Tanggung jawab profesional juga mencakup aspek metodologis, seperti melakukan praktik desain yang bertanggung jawab, yang muncul sebagai pendekatan yang mempertimbangkan aspek etika sejak awal siklus hidup, bukan sebagai renungan atau tindakan korektif. Privasi berdasarkan desain adalah salah satu contoh pendekatan ini. Prinsip ini juga menyerukan pedoman perilaku bagi individu dan konfigurasi tim AI untuk menavigasi kompleksitas etika AI secara lebih efektif. Hal ini memerlukan promosi integritas ilmiah dan pembinaan kolaborasi di antara berbagai pemangku kepentingan di sepanjang siklus hidup AI, termasuk mereka yang terdampak AI, untuk memanfaatkan beragam keahlian dalam membingkai masalah, merancang solusi, dan mengidentifikasi potensi risiko.
- **Transparansi** adalah salah satu prinsip AI yang paling banyak disebutkan. Metode yang lazim dalam AI saat ini adalah pembelajaran mesin, yang menghasilkan model dengan

"belajar" dari kumpulan data atau contoh yang sangat banyak. Misalnya, dalam kasus COMPAS, algoritma pembelajaran mesin digunakan untuk mempelajari model yang memprediksi skor risiko residivisme berdasarkan data historis dari individu yang telah melalui proses peradilan. Namun, model yang "dipelajari" ini kemungkinan besar dapat dipahami oleh manusia, sebuah karakteristik umum model yang dikembangkan melalui pembelajaran mesin, terutama yang dihasilkan melalui jaringan saraf dalam (atau pembelajaran mendalam). Inilah sebabnya mengapa model AI sering disebut buram atau kotak hitam. Keburaman ini menunjukkan bahwa manusia, termasuk pengembangnya sendiri, mungkin tidak sepenuhnya memahami bagaimana model tersebut sampai pada hasil atau prediksinya. Namun, manusia tahu bahwa model tersebut dipelajari melalui algoritma yang berupaya mengoptimalkan fungsi berdasarkan data atau contoh yang digunakan untuk proses pembelajaran yang dikenal sebagai pelatihan model. Akibatnya, gagasan transparansi menantang keburaman ini dengan bertujuan untuk membuat cara kerja dan alasan di balik keputusan model AI menjadi jelas dalam situasi tertentu. Meskipun transparansi dapat dilihat sebagai membuat kode tersedia, konseptualisasinya jauh lebih rumit. Membuka kode mungkin tidak menjelaskan fungsi model karena kompleksitasnya, dan bahkan jika bisa, jumlah orang yang akan memahaminya tetap terbatas. Dengan demikian, konsep transparansi yang bermakna telah muncul, menekankan penyediaan informasi yang relevan dan dapat ditindaklanjuti yang dapat diakses oleh para pemangku kepentingan sesuai dengan pemahaman dan kebutuhan mereka. Transparansi melampaui sekadar menyediakan informasi tentang hasil sistem AI; itu juga melibatkan menjelaskan pilihan yang dibuat di seluruh tahapan konseptualisasi, perancangan, konstruksi, evaluasi, dan penggunaan sistem AI. Bagi pengguna, transparansi melibatkan pemberitahuan ketika sistem AI membuat keputusan penting atau ketika mereka berinteraksi dengan satu, memperoleh informasi yang berguna tentang bagaimana AI membuat keputusan (berkaitan dengan konsep explainability yang kita bahas di bawah), dan mengakses informasi tentang data yang sedang digunakan. Lebih luas lagi, bagi para pemangku kepentingan seperti pembuat kebijakan, otoritas pengatur, pemimpin organisasi, atau anggota masyarakat sipil, transparansi berarti membangun dan mengelola sistem AI dengan cara yang memungkinkan pengawasan. Hal ini termasuk meningkatkan kesadaran akan kemampuan dan kendala sistem dan mendorong komunikasi terbuka di antara semua pemangku kepentingan, memberikan penjelasan untuk keputusan algoritmik, membenarkan bagaimana sistem AI beroperasi seperti merinci data yang digunakan untuk pelatihan dan bagaimana data tersebut diambil sampelnya atau diberi label, dan memastikan ketertelusuran antara keputusan ini dan hasilnya. Dengan demikian, di luar hubungan dengan keterjelasan hasil AI, transparansi juga krusial untuk memfasilitasi prinsip kendali manusia atas AI, meskipun sebagian.

- **Keterjelasan** mengacu pada kemampuan untuk memberikan wawasan atau alasan agar pengoperasian sistem AI, terutama hasilnya, menjadi jelas atau mudah dipahami oleh audiens target. Gagasan ini berbeda dari interpretabilitas model AI, yang

menunjukkan kualitas inheren model AI yang memfasilitasi pemahamannya, yang menyiratkan bahwa model berbasis aturan jauh lebih mudah diinterpretasikan daripada model yang berasal dari teknik pembelajaran mendalam. Keterjelasan, sebaliknya, juga mencakup pembuatan informasi post-hoc yang membantu menjelaskan model AI, bahkan yang dikembangkan melalui pembelajaran mendalam, dengan fokus yang disengaja pada manusia yang perlu memahami informasi ini. Upaya awal dan signifikan untuk memungkinkan keterjelasan, seperti LIME dan SHAP, terutama ditujukan kepada pengembang AI. Mereka memberikan wawasan, seperti mengidentifikasi fitur paling signifikan yang memengaruhi keputusan AI, untuk membantu pengembang menilai kesesuaian hubungan fitur tertentu untuk pengambilan keputusan, baik untuk memilih model yang tepat maupun meningkatkan model berkinerja terbaik. Baru-baru ini, perspektif ini telah diperluas untuk mencakup AI yang dapat dijelaskan yang berpusat pada manusia, yang bertujuan untuk memperluas jangkauan penerima penjelasan AI dan lebih menyeluruh menangani beragam konteks, pengetahuan sebelumnya, dan kebutuhan mereka. Kembali ke skenario COMPAS, mengadopsi perspektif yang berpusat pada manusia akan memprioritaskan kebutuhan hakim (pengguna) serta terdakwa dan perwakilan hukum mereka (mereka yang terdampak oleh keputusan AI). Pendekatan ini bertujuan untuk mengembangkan penjelasan yang memungkinkan mereka memahami proses di balik kesimpulan AI, mengevaluasi kesesuaiannya, dan memutuskan langkah selanjutnya. Langkah-langkah ini dapat bervariasi mulai dari menerima hasil dan melengkapinya dengan informasi tambahan hingga menemukan cara untuk menentang dan memperbaiki keputusan yang dimediasi AI. Oleh karena itu, kemampuan menjelaskan memainkan peran penting dalam mendorong transparansi dan memungkinkan kendali manusia atas AI. Lebih lanjut, penting untuk mengevaluasi dampak keputusan AI terhadap kemampuan menegakkan hak asasi manusia, dengan menilai apakah hal ini dicapai secara adil di berbagai kelompok sosial. Hal ini berkaitan dengan prinsip keadilan, yang akan kita bahas lebih lanjut di bawah ini.

- **Keadilan** muncul sebagai tindakan balasan terhadap bias algoritmik, sebuah istilah yang menunjukkan "penyimpangan sistematis dari kesetaraan yang muncul dalam keluaran suatu algoritma," yang dapat merugikan anggota kelompok tertentu dengan, misalnya, membatasi akses mereka terhadap tunjangan atau meningkatkan kemungkinan mereka menghadapi hukuman. ProPublica menyoroti sistem COMPAS sebagai contoh bias rasial dalam keputusan pengadilan, yang menggambarkan bagaimana tingkat positif dan negatif palsu sangat bervariasi antar ras. Sistem ini cenderung melebih-lebihkan risiko residivisme bagi terdakwa kulit hitam, yang berdampak negatif pada prospek jaminan mereka. Kesenjangan tersebut dapat ditelusuri kembali ke bias historis dalam praktik penegakan hukum, seperti peningkatan pengawasan dan tingkat penahanan di AS, tempat data pelatihan dikumpulkan. Bias AI tidak terbatas pada sistem peradilan tetapi mencakup seluruh layanan kesehatan, pendidikan, pekerjaan, pemasaran, dan banyak lagi, yang

menunjukkan bias berdasarkan ras, jenis kelamin, dan faktor-faktor lainnya. Prinsip keadilan, kemudian, menggarisbawahi keharusan untuk memerangi diskriminasi terhadap individu dan kelompok, menyoroti penghapusan bias yang tidak adil sebagai hal yang penting untuk mencegah ketidakadilan sosial dan menjaga otonomi individu. Sementara diskusi tentang keadilan dalam AI awalnya berpusat di sekitar hasil yang tidak setara, diskusi ini telah berkembang untuk mendesak para perancang dan pengguna untuk memperhatikan bias dan faktor-faktor lain yang memengaruhi diskriminasi di seluruh siklus hidup AI. menekankan pentingnya tidak hanya mencegah tetapi juga secara aktif memantau dan memitigasi bias dan diskriminasi. Namun, fokus yang signifikan diberikan pada peran penting data, yang sering kali mengandung bias sosial yang mengakar dari masa lalu atau tidak memiliki inklusivitas untuk mencerminkan keragaman mereka yang terkena dampak sistem AI. Situasi ini telah menyebabkan seruan untuk penggunaan data yang representatif dan berkualitas tinggi, dengan mempertimbangkan aspek-aspek seperti akurasi, konsistensi, dan validitas. Meskipun demikian, isu keadilan jauh melampaui cakupan data, mencakup kebutuhan untuk mengatasi bias, ketidakadilan, dan diskriminasi di seluruh perumusan masalah, pra-pemrosesan, implementasi AI, dan evaluasi hasilnya. Strategi untuk mencapai hal ini mencakup melibatkan beragam pemangku kepentingan dalam perumusan masalah dan proses desain untuk memastikan representasi, mengintegrasikan pertimbangan aksesibilitas dan melaksanakan audit algoritmik, yang berkaitan dengan prinsip akuntabilitas. Secara keseluruhan, tujuannya adalah agar setiap orang diperlakukan secara adil, tanpa diskriminasi berdasarkan ras, jenis kelamin, kebangsaan, usia, kelas sosial, keyakinan politik, agama, dan disabilitas, antara lain, dengan tujuan yang lebih luas untuk mencegah kerugian lebih lanjut bagi kelompok-kelompok yang terpinggirkan dan menghindari berlanjutnya bias historis. Selain itu, ada penekanan pada distribusi yang adil dari manfaat dan nilai yang dihasilkan oleh AI, memastikan bahwa keuntungan ini juga menjangkau mereka yang biasanya dikecualikan atau terpinggirkan.

- **Privasi** berkaitan dengan menghormati keintiman orang, kendali mereka atas informasi pribadi, dan kemampuan mereka untuk membuat pilihan tentang data mereka dan keputusan yang diambil darinya. Ini mencakup aspek-aspek seperti memperoleh persetujuan untuk menggunakan data pribadi, memungkinkan individu untuk mengontrol bagaimana data mereka digunakan, menyediakan kemampuan untuk membatasi penggunaan data, dan memberikan hak untuk memperbaiki dan menghapus data mereka. Masing-masing faktor ini memainkan peran penting dalam membentuk kembali bagaimana data dikumpulkan (dan akan) dikumpulkan dan kemudian digunakan untuk melatih model AI. Selain itu, mereka dapat memengaruhi penerapan sistem AI untuk aktivitas yang dapat melanggar privasi individu, seperti pengawasan. Prinsip ini mencakup pengembangan langkah-langkah untuk melindungi data dari akses tidak sah, yang dapat melibatkan teknik seperti enkripsi data dan privasi diferensial. Pelanggaran data yang signifikan, seperti kasus Cambridge Analytic, dan

identifikasi ulang data anonim dalam kontes Netflix untuk rekomendasi dan Sensus AS adalah contoh menonjol dari kerentanan seputar privasi data saat ini. Berbeda dengan banyak prinsip lainnya, fitur unik dalam kasus ini adalah keberadaan undang-undang yang dirancang untuk melindungi data dan privasi di sebagian besar negara di dunia (di 137 dari 194 negara pada tahun 2021). Di antara undang-undang tersebut, Peraturan Perlindungan Data Umum (GDPR) Uni Eropa memainkan peran penting dalam membingkai perdebatan tentang regulasi data, bidang ilmu data, dan ekspektasi privasi terkait AI.

- **Akuntabilitas** menggarisbawahi pentingnya membangun mekanisme untuk memastikan bahwa tanggung jawab atas hasil yang salah dan dampak buruk sistem AI ditetapkan dengan tepat. Akuntabilitas meliputi identifikasi tanggung jawab organisasi dan individu dalam menciptakan dan menerapkan sistem AI dan mengadvokasi pembentukan dan adopsi kerangka hukum untuk mendefinisikan dan menegakkan tanggung jawab ini. Perhatian khusus diberikan pada skenario di mana keputusan dibuat tanpa campur tangan manusia, di samping dampak signifikan AI terhadap keberlanjutan dan masyarakat. Akuntabilitas juga melibatkan pengaktifan auditabilitas sistem AI dan memiliki mekanisme untuk meningkatkan sistem setelah audit. Lebih lanjut, akuntabilitas menekankan penerapan strategi untuk menilai dan memitigasi risiko. Akuntabilitas meliputi pelaksanaan penilaian dampak, verifikasi bahwa sistem berfungsi dengan benar, memberikan informasi yang cukup untuk validasi pihak ketiga, membangun mekanisme untuk mengajukan banding atas hasil AI, dan menyediakan solusi untuk setiap kerugian yang mungkin ditimbulkan AI terhadap lingkungan dan individu (misalnya, kompensasi).
- **Keselamatan dan keamanan** merupakan dua dimensi penting dari setiap sistem informasi, termasuk yang didukung oleh AI. Keamanan memastikan bahwa sistem beroperasi sesuai rancangannya tanpa menyebabkan kerugian yang tidak diinginkan, sementara keamanan melibatkan kemampuan sistem untuk melindungi diri dari akses atau manipulasi yang tidak sah oleh entitas eksternal. Keamanan memerlukan konstruksi dan verifikasi sistem untuk memastikan sistem beroperasi sebagaimana mestinya dan mencegah penyalahgunaan, dengan tujuan menghilangkan risiko kerugian. Hal ini mencakup kemampuan sistem untuk berfungsi secara akurat, andal, dan tangguh, bahkan dalam kondisi yang menantang. Yang terpenting, penting untuk mempertahankan praktik pengujian dan pemantauan keamanan pasca-penerapan, mengamati perilaku sistem AI saat sistem tersebut belajar dari data baru dan beradaptasi, atau saat sistem tersebut terus bekerja dengan parameter model awal dalam berbagai skenario. Pada gilirannya, keamanan berkaitan dengan perlindungan sistem dan komponennya terhadap ancaman yang berlawanan. Keamanan bertujuan untuk memastikan sistem tetap berfungsi dan dapat diakses oleh pengguna yang sah sekaligus melindungi informasi pribadi dari pihak yang tidak berwenang.

Kategorisasi prinsip-prinsip ini, yang didasarkan pada literatur, mencakup berbagai masalah etika dalam AI. Namun, kami menyadari sifat AI yang dinamis dan terus berkembang,

yang dapat memunculkan pertimbangan dan tantangan etika baru. Oleh karena itu, kami memandang prinsip-prinsip ini bukan sebagai sesuatu yang final, melainkan sebagai titik awal untuk dialog, refleksi, dan penyesuaian yang berkelanjutan. Di bagian selanjutnya, kami akan menyajikan studi kasus untuk membahas hubungan yang saling terkait di antara prinsip-prinsip ini dan menawarkan kesempatan untuk merefleksikan perlunya perspektif sosioteknis untuk menanganinya secara lebih efektif.

BAB 2
PENDEKATAN SOSIOTEKNIS BAGI INTEGRASI ETIKA AI

2.1 APA ITU AI SOSIOTEKNIS?

Dalam buku ini, mengenali sistem AI sebagai sistem sosioteknis akan menjadi dasar untuk menganalisis permasalahan etika yang kita hadapi dalam AI dan TI. Dalam bab ini, kami menjelaskan mengapa kami mengadopsi perspektif sosioteknis untuk memahami dan mempelajari AI, serta memberikan contoh perspektif terkini yang merangkulnya untuk mengintegrasikan isu-isu etika utama ke dalam pengembangan proyek AI.

Pendekatan ini tidak hanya mempertimbangkan aspek teknis pengembangan teknologi ini, tetapi juga pengaruh sosial yang kompleks yang diperlukan untuk merancang, mengembangkan, mengimplementasikan, dan menggunakan AI. Sederhananya, menganggap AI sosioteknis berarti mengakui bahwa AI tidak beroperasi dalam ruang hampa, melainkan dalam konteks kompleks yang menuntut perhatian pada berbagai aspek.

Gagasan "sistem sosioteknis" muncul sebelum kaitannya dengan AI. Para peneliti telah mengkarakterisasi sistem sosioteknis dengan berbagai cara. Baxter dan Sommerville, misalnya, membahas metode perancangan sistem sosioteknis sebagai pendekatan yang mempertimbangkan “faktor manusia, sosial, dan organisasi, serta faktor teknis dalam perancangan sistem organisasi. Hasil penerapan metode ini adalah pemahaman yang lebih baik tentang bagaimana faktor manusia, sosial, dan organisasi memengaruhi cara kerja dilakukan dan sistem teknis digunakan”. Namun, dasar definisi ini dapat ditelusuri kembali ke tahun 1950-an.

Di Tavistock Institute di London, konsep “sistem sosioteknis” muncul sebagai respons terhadap beberapa proyek yang mengembangkan industri pertambangan batu bara Inggris. Skenario pascaperang berarti bahwa teknologi baru sedang diimplementasikan, tetapi juga bahwa hubungan interpersonal untuk mengelola dinamika ketenagakerjaan baru perlu ditangani, termasuk pengaturan organisasi untuk mencapai pengembangan industri dan produktivitas yang komprehensif.

Singkatnya, usulan utama dari "paradigma baru" ini adalah pergeseran dalam memahami dan merancang organisasi kerja, yang menekankan transisi dari pendekatan teknokratis murni menuju pengakuan dan pemberian nilai pada pentingnya memahami peran manusia dalam konteks kelembagaan. Oleh karena itu, paradigma lama tentang keharusan teknologi, yang menempatkan mesin sebagai pusat, dan manusia dianggap sekadar pelengkap, tergantikan dengan gagasan bahwa manusia dan dimensi sosial lainnya saling terhubung dengan teknologi.

Dengan demikian, penelitian sosioteknis telah dikarakterisasikan sebagai upaya yang didasarkan pada manfaat dua arah yang diperoleh dari persinggungan elemen sosial dan teknis. Persinggungan tersebut membutuhkan resiprositas antara teknologi dan masyarakat, yang membentuk bagaimana kedua dimensi tersebut berkembang. Dengan kata lain, keberhasilan sistem sosioteknis dapat dievaluasi berdasarkan keberhasilan interaksi tersebut.

Seiring berkembangnya gagasan "sistem sosioteknis", demikian pula kekhususannya yang memungkinkan kita untuk membuat hubungan yang lebih mendalam antara aspek sosial dan teknis dari teknologi, tidak hanya berdasarkan pada tingkat organisasi atau manajemen. Pinch dan Bijker, misalnya, mengikuti kerangka kerja SCOT (Konstruksi Sosial Teknologi), yang menyatakan bahwa pengembangan teknologi merupakan proses teknis sekaligus sosial, dengan menyatakan bahwa artefak dan sistem teknologi dibentuk oleh faktor-faktor sosial, ekonomi, dan budaya, yang menyiratkan bahwa teknologi tidak berevolusi terutama melalui inovasi para insinyur individu atau perangkat teknologi tertentu.

Lebih spesifik lagi, Bijker memperkenalkan konsep-konsep baru yang menambah kedalaman pandangan sosioteknis. Salah satu contohnya adalah gagasan fleksibilitas interpretatif. Konsep ini menyatakan bahwa artefak teknologi dapat memiliki makna dan penggunaan yang berbeda bagi kelompok sosial yang berbeda. Ini berarti bahwa desain, pengembangan, dan penggunaan teknologi terbuka untuk interpretasi berdasarkan kebutuhan, nilai, dan konteks sosial pengguna. Fleksibilitas ini memungkinkan solusi teknologi yang berbeda untuk masalah yang sama. Yang ditunjukkan oleh Bijker adalah bahwa makna yang dikaitkan dengan suatu teknologi atau artefak tidak didasarkan pada teknologi itu sendiri. Lebih lanjut, perbedaan ini memiliki hubungan yang lebih mendalam dengan apa yang terkandung dalam sistem sosioteknis: "kelompok sosial yang relevan tidak hanya melihat aspek-aspek berbeda dari satu artefak. Makna yang diberikan oleh kelompok sosial yang relevan sebenarnya membentuk artefak tersebut". Oleh karena itu, kita tidak dapat benar-benar berevolusi dan mengembangkan kemajuan teknologi tanpa pemahaman sosioteknisnya.

Konsep sistem sosioteknis mengacu pada sistem yang bergantung pada kombinasi infrastruktur teknis, perilaku manusia, dan institusi sosial agar berfungsi secara efektif. Lebih spesifik lagi, van de Poel mengusulkan bahwa sistem tersebut terdiri dari tiga komponen fundamental:

1. Artefak teknis, yang merupakan objek fisik yang dirancang untuk fungsi teknis tertentu, yaitu memiliki kehadiran fisik,
2. Agen manusia, individu yang melakukan tindakan intensional dan berinteraksi dengan artefak teknis, dan
3. Institusi, yang merupakan aturan atau norma sosial yang mengatur perilaku, yang menetapkan ekspektasi untuk perilaku moral.

Selain ketiga elemen ini, sistem AI memiliki dua komponen unik tambahan yang memengaruhi pemahaman sosioteknisnya. Komponen-komponen tersebut adalah:

1. Agen buatan, dan
2. Norma teknis.

Sementara agen manusia dan institusional biasanya dipahami dalam konteks intensionalitas, agen buatan dan norma teknis, menurut van de Poel, beroperasi dalam kerangka kerja yang ditentukan oleh mekanisme kausal atau fisik. Agen buatan berbeda dari artefak teknis konvensional karena kemampuannya untuk menunjukkan atau meniru karakteristik mirip manusia seperti otonomi, interaktivitas, dan adaptivitas .

Tidak seperti artefak teknis tradisional, yang utamanya bergantung pada struktur fisik, agen buatan memiliki kemampuan untuk beradaptasi dengan berbagai konteks dan berinteraksi dengan agen lain (buatan atau bukan), sehingga membentuk perilaku yang lebih kompleks dan dinamis. Namun, seperti yang kita ketahui, meskipun agen buatan dapat menunjukkan karakteristik yang lebih dinamis daripada artefak teknis, kemampuannya tidak memiliki sifat-sifat inheren manusia seperti kesadaran, kehendak bebas, emosi, dan otonomi moral.

Norma teknis bergantung pada keberadaan agen buatan yang tidak mengikuti aturan sosial tradisional, karena aturan-aturan ini berkaitan dengan interaksi manusia. Sebaliknya, norma teknis berfungsi sebagai padanan AI untuk institusi, menyediakan seperangkat aturan dan pedoman yang mengatur perilaku agen buatan. Norma teknis dipahami secara kausal, memberikan batasan operasional di mana agen buatan dapat berfungsi, memastikan konsistensi, dan menetapkan batasan bagi agen buatan untuk beroperasi, yang memungkinkan mereka berfungsi dalam sistem AI tanpa memiliki intensionalitas mirip manusia.

Oleh karena itu, AI, yang dipahami sebagai sistem sosioteknis, berarti mengakui adanya jaringan kompleks yang melibatkan interaksi antara elemen-elemen teknis, seperti algoritma, agen buatan, dan norma-norma teknis, dengan elemen-elemen sosial, seperti perilaku manusia dan aspek kelembagaan, termasuk norma-norma budaya dan kerangka regulasi. Efektivitas dan dampak suatu sistem dibentuk oleh interaksi-interaksi ini, yang beroperasi dalam siklus umpan balik dinamis di mana teknologi memengaruhi, dan dipengaruhi oleh, praktik-praktik manusia dan masyarakat, termasuk pertimbangan etika dan pedoman hukum.

Mengadopsi perspektif sosioteknis terhadap AI menyoroti perlunya memahami sistem AI dalam konteks yang lebih luas, tidak hanya mempertimbangkan komponen teknisnya tetapi juga konsekuensi sosialnya, yang memiliki implikasi signifikan bagi pengembangan etika AI, yaitu merancang sistem AI yang selaras dengan nilai-nilai sosial, sehingga memastikan bahwa teknologi memberikan manfaat yang lebih besar. Untuk mencapai tugas yang rumit seperti itu, mengadopsi pandangan sosioteknis terhadap AI memerlukan pendekatan kolaboratif yang melibatkan ahli etika, insinyur, pembuat kebijakan, dan ilmuwan sosial, di antara para profesional lainnya, untuk memastikan bahwa sistem AI dirancang dan diatur dengan cara yang mendukung prinsip-prinsip etika dan nilai-nilai sosial.

2.2 ELEMEN KUNCI PENDEKATAN SOSIOTEKNIS TERHADAP AI

Setelah Anda memahami arti dari perspektif sosioteknis, kami harap Anda dapat memahami mengapa kami yakin hal ini penting untuk memajukan etika AI. Cara sudut pandang sosioteknis dapat memengaruhi etika AI beragam, dan di sini, kami akan memberikan beberapa contoh yang menyoroti elemen kunci pendekatan sosioteknis agar Anda dapat mengeksplorasi berbagai cara perspektif ini dapat bermanfaat bagi pekerjaan Anda di bidang AI.

Desain AI Berbasis Nilai

Artefak teknis dapat mewujudkan nilai-nilai jika dirancang dengan tujuan spesifik yang selaras dengan nilai-nilai tersebut dan jika penggunaannya kondusif untuk mencapainya. Misalnya, jika artefak teknis dirancang dengan fitur keselamatan, dan fitur-fitur tersebut berkontribusi pada keselamatan jika digunakan dengan benar, maka artefak tersebut mewujudkan nilai keselamatan. Konsep penanaman nilai dalam artefak teknis berpusat pada gagasan bahwa pilihan desain mencerminkan nilai-nilai tertentu yang juga dapat dipahami sebagai pilihan etis.

Selain itu, pilihan-pilihan ini memengaruhi bagaimana artefak berfungsi dalam praktik. Hal yang sama terjadi pada institusi yang mewujudkan nilai-nilai melalui aturan formal dan standar praktik, yang mempromosikan hasil dan tujuan spesifik. Namun, sebagaimana telah disebutkan, agen buatan tidak mengikuti serangkaian aturan ini. Oleh karena itu, menerjemahkan nilai-nilai ini ke dalam norma teknisnya dapat memungkinkan kita untuk memandu perilaku AI yang otonom dan adaptif.

Penerjemahan nilai-nilai ini ke dalam norma teknis bukan sekadar fitur pelengkap, melainkan aspek fundamental dari desain AI yang etis, krusial untuk memastikan bahwa teknologi AI berfungsi dengan cara yang bermanfaat dan selaras dengan prinsip-prinsip etika. Memasukkan pertimbangan etika ke dalam desain AI, yang disebut sebagai desain etis, mengamanatkan bahwa sistem AI dirancang tidak hanya dengan mempertimbangkan tujuan fungsional spesifik, tetapi juga dengan kesadaran yang tajam akan dampak sosialnya yang lebih luas.

Misalnya, AI yang bertugas memoderasi konten di platform digital harus menyeimbangkan efikasi operasionalnya dengan pertimbangan keadilan dan kebebasan berekspresi. Keseimbangan ini dapat memastikan bahwa moderasi tidak secara tidak adil menargetkan kelompok tertentu atau membungkam wacana yang sah, sehingga mewujudkan nilai-nilai etika keadilan dan penghormatan terhadap hak-hak individu. Secara keseluruhan, norma teknis berbeda dari norma sosial karena norma tersebut ditegakkan melalui kode dan algoritma, alih-alih ekspektasi sosial atau moral. Oleh karena itu, tujuannya adalah untuk memahami bagaimana keputusan manusia tentang fitur-fitur teknis ini memengaruhi konteks sosioteknis AI.

Lebih lanjut, aspek kunci dari perspektif sosioteknis adalah mengakui bahwa nilai-nilai dapat berkembang seiring sistem digunakan dan diadaptasi dari waktu ke waktu. Bukan hanya desain awal yang penting; proses perancangan ulang dan umpan balik yang berkelanjutan sangat penting untuk memastikan bahwa sistem terus mewujudkan nilai-nilai yang diinginkan seiring perkembangan dan adaptasinya. Oleh karena itu, pemantauan berkelanjutan diperlukan untuk menjaga keselarasan dengan prinsip-prinsip etika dan nilai-nilai sosial, sebagaimana kita belajar dari interaksi dan simbiosis sosioteknis AI dan masyarakat.

Mekanisme umpan balik sangat penting untuk menyelaraskan sistem AI dengan nilai-nilai sosial. Mekanisme ini memungkinkan sistem AI untuk beradaptasi berdasarkan masukan pengguna dan pemangku kepentingan, yang mencerminkan komitmen berkelanjutan terhadap keselarasan etika seiring perkembangan nilai-nilai sosial. Proses penyesuaian dinamis ini penting karena memungkinkan sistem AI untuk tetap relevan dan selaras secara

etika seiring waktu meskipun ada perubahan dalam norma dan nilai-nilai sosial. Konsep mesin moral dapat menggambarkan hal ini.

Misalnya, kendaraan otonom dapat diprogram untuk mengambil keputusan dalam situasi yang sarat moral, seperti skenario kecelakaan di mana kerugian tidak dapat sepenuhnya dihindari tetapi dapat diminimalkan. Proses pengambilan keputusan dalam kasus-kasus seperti itu sarat dengan pertimbangan etika, dengan mempertimbangkan faktor-faktor seperti keselamatan penumpang versus pejalan kaki. Demikian pula, dalam layanan kesehatan, sistem AI seperti robot di ruang gawat darurat mungkin perlu memprioritaskan perawatan pasien berdasarkan tingkat keparahan dan urgensi, dengan mengintegrasikan nilai-nilai keadilan dan kesetaraan dalam prioritas perawatan medis.

Adaptabilitas: Tata Kelola Etika yang Tangguh

Chopra dan Singh berpendapat bahwa pengambilan keputusan moral dalam AI harus dibedakan dari perspektif sosioteknis. Para penulis mengkritik bahwa pendekatan tradisional terhadap pengambilan keputusan moral dalam AI seringkali bergantung pada perspektif teoretis keputusan, yang berarti pendekatan tersebut hanya berfokus pada agen individu atau permasalahan etika individu. Mereka mengklaim bahwa pendekatan ini mungkin terlalu sempit karena pengambilan keputusan moral sangat dipengaruhi oleh konteks. Oleh karena itu, sebaliknya, mereka berpendapat bahwa etika perlu dipertimbangkan dari perspektif sosioteknis yang lebih luas untuk mencapai analisis yang lebih holistik.

Untuk mengembangkan pendekatan mereka, mereka mempertimbangkan tiga elemen sistem sosioteknis: (1) pemangku kepentingan; (2) informasi, yaitu nilai-nilai pemangku kepentingan, norma-norma preskriptif, dan hasil; dan (3) proses tata kelola, termasuk untuk tujuan respesifikasi sistem sosioteknis. Respecifying mengacu pada revisi atau pendefinisian ulang parameter suatu sistem, model, atau proyek, dan hal ini penting karena melibatkan aktivitas interaktif di antara para pemangku kepentingan untuk memastikan bahwa norma-norma sistem sosioteknis selaras dengan pertimbangan etika dari waktu ke waktu.

Untuk ini, mereka menetapkan tiga aktivitas tata kelola: desain, penetapan, dan adaptasi. memastikan bahwa tata kelola etika AI mewakili dinamisme yang diberikan oleh lensa sosioteknis. Desain, bagi Chopra dan Singh, melibatkan penciptaan sistem sosioteknis yang memenuhi persyaratan para pemangku kepentingan, dengan mempertimbangkan nilai-nilai mereka dan terlibat dengan mereka selama proses berlangsung. Penetapan mengacu pada perilaku dalam sistem sosioteknis, memastikan bahwa mereka bertindak seperti yang diharapkan. Adaptasi menjawab kebutuhan untuk penyesuaian berkelanjutan pada sistem berdasarkan hasil dan perubahan persyaratan. Pendekatan tata kelola ini mendorong sistem yang fleksibel dan responsif yang dapat berkembang seiring konteks.

Adaptabilitas, tema sentral dari perspektif sosioteknis, mencerminkan kebutuhan sistem untuk berkembang sebagai respons terhadap informasi baru dan perubahan norma. Para peneliti menyoroti bahwa norma mungkin perlu diubah untuk mengarahkan hasil ke arah yang diinginkan. Selain itu, mekanisme komputasi, seperti algoritma, mungkin perlu beradaptasi untuk memenuhi standar baru. Adaptasi ini mendorong inovasi dan memungkinkan perbaikan berkelanjutan dalam sistem sosioteknis.

Implikasi dari penerapan pandangan sosioteknis terhadap AI mengubah pola intervensi yang terisolasi atau individual serta analisis masalah etika dan, sebagai gantinya, berfokus pada kerangka kerja yang lebih luas yang didasarkan pada pertimbangan konteks di mana AI beroperasi, dengan fokus pada tata kelola dan adaptasi untuk memastikan hasil yang etis. Oleh karena itu, untuk mengembangkan sistem AI yang etis, kita tidak hanya merespons keputusan individual, tetapi juga mengeksplorasi bagaimana keputusan tersebut berinteraksi dengan berbagai elemen dan pemangku kepentingan dalam ekosistem AI untuk memastikan tata kelola yang lebih kuat.

Interdependensi: diperlukan kerja multidisiplin dan interdisipliner

Interdependensi antara teknologi dan masyarakat merupakan inti dari pembahasan AI sebagai sistem sosioteknis. Mitos dan narasi teknologi berdampak signifikan terhadap adopsi dan penggunaan teknologi AI, memengaruhi cara pengembang, warga negara, dan pembuat kebijakan memandang dan berinteraksi dengan AI, sebagaimana dicatat oleh Sartori dan Theodorou. Kedua penulis menekankan pentingnya meningkatkan kesadaran masyarakat, yang memungkinkan mereka mengadopsi teknologi baru secara kritis.

Pertimbangan untuk kesadaran ini menggarisbawahi persyaratan lain: pendekatan multidisiplin untuk menggabungkan perspektif dari berbagai pemangku kepentingan. Bagi Sartori dan Theodorou, mengadopsi pandangan sosioteknis tentang sistem AI mengharuskan kita untuk "melibatkan semua peserta dalam proses konstruksi dalam pendekatan ko-kreasi" . Narasi yang membentuk realitas dan konteks saat ini di mana AI diciptakan mencerminkan persepsi dan keyakinan yang memengaruhi masa depan AI.

Bagaimana narasi ini dapat berubah lintas budaya sangatlah relevan, karena hal ini dapat diterjemahkan langsung ke dalam prinsip-prinsip etika yang kita harapkan dari AI dan keputusan yang kita buat tentang cara mengembangkannya. Dengan demikian, pentingnya pekerjaan interdisipliner dalam AI disorot oleh kebutuhan untuk memahami dan mengatasi saling ketergantungan yang kompleks antara elemen sosial dan teknis, yang didukung oleh pandangan sosioteknis.

Dalam praktiknya, hal ini dapat diterjemahkan ke dalam cara-cara spesifik untuk mengintegrasikan etika dengan mengadopsi pandangan sosioteknis yang memiliki konteks disiplin yang memediasi saling ketergantungan faktor-faktor teknis dan sosial. Misalnya, ketika mengintegrasikan etika ke dalam pendidikan Ilmu Komputer, Goetze berpendapat bahwa pendekatan interdisipliner lebih efektif daripada pendekatan multidisiplin karena pendekatan tersebut menghubungkan erat konten etika dengan pembelajaran teknis, yang memungkinkan pemahaman yang lebih dalam tentang isu-isu etika dalam komputasi. Penulis mengakui bahwa pendidikan teknis tidak cukup tanpa menggabungkan pertimbangan etika. Mereka menyarankan bahwa siswa harus mampu menganalisis dan membangun argumen etika serta memahami dampak sosial dari teknologi komputasi.

Untuk mencapai hal ini, mereka perlu dibenamkan dalam konteks dan implikasi etika dari pekerjaan mereka, yang juga dapat memperoleh manfaat dari pendekatan transdisipliner. Integrasi yang mendalam antara konsep etika dan teknis berpotensi mengarah pada paradigma baru dalam komputasi, dengan keberhasilan etika sama pentingnya dengan

keberhasilan teknis. di mana mereka berargumen untuk menanamkan keadilan algoritmik dalam pandangan sosioteknis tentang sistem informasi justru untuk mengakui dan secara aktif memasukkan saling ketergantungan antara aspek sosial dan teknis, karena bagi mereka:

> *"Tanpa perspektif yang koheren yang mengakui saling ketergantungan antara aspek sosial dan teknis AI, organisasi mungkin enggan untuk mengatasi masalah ini secara efektif. Jika mereka memperlakukan (ketidak)adilan algoritmik sebagai masalah teknis semata, mereka mungkin berasumsi bahwa menambahkan elemen sosial akan cukup menyelesaikan ketidakadilan."*

Dari sudut pandang sosioteknis, para penulis menekankan bahwa keadilan dalam AI lebih dari sekadar mengurangi bias teknis atau mencapai hasil yang adil. Hal ini melibatkan pertimbangan bagaimana struktur sosial, interaksi manusia, dan sistem teknis secara kolektif memengaruhi hasil algoritmik. Pendekatan teknis semata dapat mengabaikan faktor-faktor penting seperti nuansa budaya, norma sosial, dan konteks historis, yang menyebabkan hasil yang tidak adil meskipun ada upaya teknis untuk memastikan keadilan. Dengan demikian, pendekatan sosioteknis membawa kita menjauh dari pendekatan (teknis) tradisional, sebagaimana menyebutnya, yang seringkali berfokus pada keadilan distributif, mereduksi keadilan menjadi masalah statistik. Pendekatan sosioteknis mempertimbangkan aspek keadilan lainnya, seperti keadilan interaksional dan prosedural, yang krusial dalam memastikan bahwa sistem AI secara teknis baik dan adil secara sosial.

Situasi: Konteks yang Mewujud

Memahami dampak AI yang lebih luas terhadap masyarakat sangatlah penting, dan untuk mencapainya, para peneliti telah menggunakan konsep-konsep krusial dari para filsuf feminis untuk menekankan persyaratan kontekstual dari sudut pandang sosioteknis. Di awal diskusi tentang pengintegrasian pandangan sosioteknis dalam AI, menyajikan pendekatan "algoritma situasi".

Mereka berpendapat bahwa algoritma, yang seringkali dipandang sebagai entitas teknis belaka, semakin banyak diteliti sebagai sistem sosioteknis dengan implikasi terhadap ketimpangan sosial dan hierarki budaya. Perspektif sosioteknis, mereka tekankan, menghindari pandangan simplistis yang memisahkan algoritma dari konteks manusianya dan mempertimbangkan bagaimana sistem ini berinteraksi dengan tatanan sosial kita. Dengan demikian, hal ini sejalan dengan gagasan "budaya algoritmik", di mana algoritma memainkan peran penting dalam memilah dan mengklasifikasikan, yang pada akhirnya memengaruhi lanskap budaya kita.

Oleh karena itu, para penulis menggunakan kritik Harding terhadap bias dan struktur kekuasaan yang secara historis membentuk pengetahuan ilmiah sebagai titik awal. Menurut Harding, memahami produksi pengetahuan membutuhkan pengakuan perspektif siapa yang terwakili dan siapa yang diuntungkan darinya. Pendekatan ini menantang gagasan objektivitas "netral nilai", dengan menyatakan bahwa semua produksi pengetahuan dipengaruhi oleh konteks sosial politik yang lebih luas.

Dalam konteks AI, kegagalan pengenalan wajah untuk mengenali wajah berwarna secara akurat, misalnya, mencerminkan pola bias rasial yang lebih dalam yang tertanam dalam sistem teknologi yang berakar pada representasi yang tidak setara dalam set data pelatihan. Sebaliknya, konsep Harding tentang "objektivitas yang kuat", yang digunakan oleh Draude, mengusulkan bahwa untuk menangkal ketidakseimbangan kekuasaan dalam produksi pengetahuan, penelitian harus dimulai dari perspektif mereka yang paling terdampak oleh ketidaksetaraan. Bentuk objektivitas yang lebih kuat ini mengakui posisinya dan berusaha untuk menginterogasi struktur kekuasaan yang ada.

Upaya penelitian, seperti pengembangan AI yang mengadopsi objektivitas yang kuat dimulai dengan mempertanyakan peran mereka sendiri dalam sistem ketidaksetaraan dan mempertimbangkan kelompok mana yang mungkin diuntungkan atau dirugikan. Pendekatan ini selaras dengan gagasan untuk menempatkan sistem algoritmik dalam konteks sosiopolitiknya, memastikan bahwa proses desain memperhitungkan keterikatan algoritma dalam hierarki kekuasaan dan kerangka budaya yang ada.

Secara keseluruhan, berpendapat bahwa "untuk menghasilkan solusi sosioteknis yang kurang bias dan lebih akuntabel, sangat penting untuk menempatkan sistem algoritmik dan proses desainnya, yaitu untuk memahami dan mengatasi keterikatannya dalam konteks politik, sosiokultural, dan struktur kekuasaan yang ada." Dengan demikian, untuk menciptakan sistem algoritmik yang akuntabel dan kurang bias, mereka mengusulkan pendekatan sistemik yang mengintegrasikan kerangka kerja "4P": People (Orang), Place (Tempat), Power (Kekuasaan), dan Participation (Partisipasi). Kerangka kerja ini membahas dampak yang lebih luas dari sistem algoritmik dengan berfokus pada pertanyaan-pertanyaan kunci tentang siapa yang terlibat, siapa yang diuntungkan, dan siapa yang mungkin terdampak negatif.

2.3 CONTOH PENDEKATAN SOSIOTEKNIS TERHADAP AI

Pada bagian ini, kami ingin menyoroti dua pendekatan yang telah merangkul perspektif sosioteknis dan elemen-elemen utamanya. Contoh-contoh ini membantu kita memvisualisasikan bagaimana pertimbangan etika dapat diintegrasikan ke dalam proses pengembangan AI.

Feminisme Data

Sehubungan dengan situasionalitas, D'Ignazio dan Klein mengusulkan Feminisme Data untuk menekankan pengalaman hidup sebagai sumber penting untuk memahami dan membentuk penggunaan serta keterbatasan data. Usulan ini mengajak kita untuk memahami data, input utama pembelajaran mesin, sebagai representasi reduktif yang niscaya dari pengalaman beberapa orang, di mana kehidupan diterjemahkan menjadi angka, kata, atau gambar tertentu, sehingga banyak bagian dari pengalaman hidup mereka tidak diperhitungkan.

Pada saat yang sama, feminisme data mengajak kita untuk memperhatikan bahwa orang lain, biasanya yang lebih istimewa, memainkan peran berpengaruh dalam menghitung, memberi label, menganalisis, menggunakan, dan memanfaatkan nilai data. Kedua aktor, baik yang datanya digunakan maupun yang menggunakan data tersebut, merupakan bagian dari

konteks di mana berbagai struktur penindasan (yaitu, berdasarkan gender, ras, kelas, kemampuan, usia, seksualitas, geografi, dan lainnya) hidup berdampingan dan berlipat ganda. Oleh karena itu, feminisme data menyoroti perlunya mengkaji konteks di mana data dihasilkan serta tujuan penggunaannya dari sudut pandang interseksional.

Interseksionalitas memberikan perhatian khusus pada beban gabungan dari pengalaman orang-orang yang hidup di bawah berbagai tingkat penindasan, seperti perempuan kulit berwarna atau migran penyandang disabilitas dari negara-negara berkembang yang seringkali luput dari perhatian orang-orang yang tidak menanggung beban tersebut. Feminisme data menekankan risiko yang berasal dari kesenjangan yang signifikan antara pengalaman hidup mereka yang membuat keputusan tentang sistem berbasis data, seringkali individu dari kelompok sosial yang homogen dan istimewa (misalnya, kulit putih, laki-laki, cisgender, nondisabilitas, berpendidikan di negara-negara berkembang) dan pengalaman mereka yang terdampak oleh sistem ini, yang biasanya merupakan populasi yang lebih beragam. Mengakui kesenjangan ini merupakan titik awal yang penting untuk mengenali tantangan dalam mengidentifikasi bagaimana sistem berbasis data dapat menyebabkan kerugian atau melestarikan bias.

Mengambil sudut pandang interseksional juga dapat membantu kita menghindari salah tafsir data sebagai artefak yang objektif dan netral. Sebaliknya, hal ini mendorong kita untuk memahami data sebagai produk dari hubungan sosial yang tidak setara, di mana berbagai aktor dan elemen, seperti manusia, termasuk tujuan dan praktik mereka, teknologi yang tersedia dan norma kelembagaan dapat memengaruhi karakteristik data. Misalnya, dalam kasus COMPAS, feminisme data mengajak kita untuk mempertanyakan mengapa terdapat lebih banyak data tentang terdakwa berkulit hitam daripada berkulit putih dan bagaimana hal itu berkaitan dengan pengawasan dan patroli yang secara historis lebih luas di lingkungan yang sebagian besar dihuni oleh orang kulit berwarna.

Oleh karena itu, penggunaan data yang kita kenali sangat bias memerlukan refleksi atas validitas dan keterbatasannya, mempertanyakan seberapa baik data tersebut merepresentasikan karakteristik pengalaman orang yang ingin kita ukur, dan memutuskan apakah dan bagaimana menggunakannya sebagai masukan untuk proyek AI kita. Tujuan menyoroti pengalaman orang-orang sebagai sumber data dan pengambil keputusan seputar data adalah untuk membantu kita menelaah bagaimana ketidakseimbangan kekuatan yang ada memengaruhi jenis teknologi berbasis data yang dikembangkan dan diterapkan secara global. Di sini, D'Ignazio dan Klein mengusulkan untuk mengkaji tujuan siapa yang diprioritaskan dalam proyek berbasis data, siapa yang diuntungkan dari proyek-proyek ini, dan kehidupan siapa yang "didatakan" dan dipengaruhi oleh hasilnya. Ini termasuk mempertimbangkan pekerjaan moderasi dan pelabelan data yang tersembunyi dan dibayar rendah, serta dampak lingkungan dari ekstraksi sumber daya alam yang diperlukan untuk mempertahankan infrastruktur data dan AI. Dalam hal ini, feminisme data berusaha mempertanyakan apakah sistem berbasis data berkontribusi untuk mempertahankan status quo atau memperkuat dan bahkan meningkatkan ketidakseimbangan kekuatan yang ada.

Lebih lanjut, feminisme data mendorong kita untuk menggunakan data dan AI dengan tujuan baru: "menantang kekuatan." Ini berarti menggunakan data untuk menumbangkan asimetri kekuatan yang ada, dengan berupaya menghormati agensi komunitas yang rentan. Gagasan kunci dalam hal ini adalah merangkul pluralisme untuk mengembangkan sistem berbasis data. Setelah mengakui reduksionisme data kuantitatif, feminisme data merekomendasikan penggabungan sistem berbasis data dengan proses inklusif dan partisipatif untuk menginformasikan pengembangan dan implementasinya dengan wawasan berbasis bukti dari berbagai perspektif lokal.

Pendekatan pluralistik ini, yang terkait dengan dimensi interdependensi yang dibahas sebelumnya, harus dilengkapi dengan mekanisme transfer pengetahuan kepada dan dari masyarakat terdampak, serta pembangunan infrastruktur sosial yang dibutuhkan untuk mendukung intervensi berbasis data teknis, sehingga memastikan adaptabilitas jangka panjang. Prinsip lain dari feminisme data termasuk menghargai emosi dan perwujudan melalui penggambaran data dan visualisasi yang netral, mengevaluasi kembali biner dan hierarki dalam data yang tersedia, dan membuat kerja di balik sistem berbasis data terlihat dan dihargai.

Keadilan Desain

Design Justice karya Costanza-Chock menggunakan konsep situasional untuk mentransformasi desain solusi berbasis teknologi, sebuah tugas yang lebih luas daripada sekadar menangani permasalahan spesifik seputar data atau AI. Penulis mengkritik bagaimana nilai, praktik, narasi, lokasi, dan pedagogi yang berlaku dalam desain mereproduksi ketimpangan sistematis, mempertahankan distribusi kekuasaan yang ada dengan, misalnya, menyesuaikan kemampuan teknologi untuk kelompok sosial yang lebih dominan dan menguntungkan, dan mengabaikan kebutuhan kaum minoritas, sehingga menciptakan beban tambahan bagi mereka.

Karya ini juga mengkaji bagaimana praktik dan narasi desainer menyembunyikan dan terkadang menyalahgunakan kontribusi pengguna, menghapus perubahan kolektif dan kumulatif yang mengarah pada terobosan mereka. Dengan demikian, Design Justice membantu kita menganalisis peran desain dalam melanggengkan ketimpangan sosial dan mengajak kita untuk mengupayakan distribusi manfaat dan beban desain yang lebih adil, dengan tujuan mencapai keadilan sosial.

Menggemakan tema-tema dari Data Feminism, Design Justice menyoroti kesenjangan antara mereka yang merancang teknologi dan komunitas yang ingin mereka layani. Mengacu pada aktivisme disabilitas dan mantranya, "tidak ada tentang kita tanpa kita," penulis memberikan argumen kuat yang mendukung keterlibatan anggota masyarakat yang secara langsung terdampak oleh teknologi dalam proses desain. Yang terpenting, keterlibatan ini harus substantif dan berkelanjutan di sepanjang siklus hidup desain, sejak awal. Berbeda dengan pendekatan desain partisipatif lainnya, Design Justice mengadvokasi partisipasi aktif anggota masyarakat dalam mendefinisikan ruang lingkup proyek dan membingkai permasalahan, memastikan bahwa pengalaman hidup mereka secara signifikan memengaruhi

definisi isu-isu yang akan ditangani oleh proses desain. Ini berarti partisipasi lebih dari sekadar bertukar pikiran dan membantu tim desain dalam pengujian prototipe.

Partisipasi juga melibatkan partisipasi dalam menentukan permasalahan mana yang perlu ditangani dan bagaimana cara mengatasinya. Pendekatan ini didukung atas dasar keadilan, tetapi juga memiliki manfaat praktis: wawasan unik, pengalaman hidup, dan pengetahuan tersirat dari anggota masyarakat dapat menghasilkan ide dan perspektif inovatif yang mungkin tidak ditemukan oleh pihak di luar masyarakat. Melibatkan anggota komunitas dalam merumuskan masalah juga dapat mengalihkan kita dari solusionisme teknosentris ketika menangani isu-isu sosial yang kompleks dan bernuansa.

Design Justice juga mengadvokasi tim desain yang beragam yang mencakup anggota komunitas, tetapi juga berupaya mengubah metode desain yang secara tradisional ekstraktif menjadi mekanisme yang memungkinkan komunitas menerima penghargaan, visibilitas, keuntungan, dan kepemilikan atas artefak yang dirancang sebagai imbalan atas kontribusi mereka terhadap proses desain. Transformasi ini tidak hanya mempromosikan keadilan sebagai nilai desain terintegrasi, tetapi juga dapat membantu memastikan adaptabilitas jangka panjang dari teknologi yang dirancang.

Lebih lanjut, Design Justice menekankan pentingnya mengintegrasikan interseksionalitas ke dalam artefak yang digunakan untuk merancang dan menilai teknologi. Constanza-Chock berpendapat bahwa kita “perlu mengembangkan kisah pengguna yang interseksional, pendekatan pengujian, data pelatihan, tolok ukur, standar, proses validasi, dan penilaian dampak, di antara banyak alat lainnya”. Dengan demikian, Feminisme Data dan Keadilan Desain menawarkan serangkaian wawasan dan saran praktis yang mengoperasionalkan beberapa aspek perspektif sosioteknis untuk menganalisis dan membangun proyek AI, mengintegrasikan isu-isu etika yang terkontekstualisasi ke dalam praktik kerja dan desain data. Sebelum beralih ke cara-cara menangani isu-isu etika dalam tahapan-tahapan spesifik proyek AI, bab selanjutnya akan membahas isu-isu etika spesifik yang masih agak tersembunyi dari perdebatan umum dalam Etika AI dan membutuhkan perhatian lebih seiring kita mengembangkan bidang ini lebih lanjut.

BAB 3
KEBERLANJUTAN DAN KRISIS REPLIKASI

3.1 KEBERLANJUTAN

Pada bab-bab sebelumnya, kami memperkenalkan dan memberikan contoh prinsip-prinsip inti dalam etika AI serta pendekatan sosioteknis untuk melengkapinya dan mengatasi berbagai permasalahan etika yang muncul dari perancangan, penggunaan, dan implementasi sistem AI. Topik-topik ini dapat dikategorikan dalam diskusi "arus utama" di bidang Etika AI. Dalam bab ini, kami ingin mengkaji dua isu yang telah keluar dari arus utama dan baru-baru ini mendapat perhatian: keberlanjutan dan krisis replikasi. Kami sangat yakin bahwa topik-topik ini membutuhkan perhatian lebih lanjut seiring kami terus mengembangkan bidang etika AI.

Van Wynsberghe dapat diakui sebagai salah satu peneliti pertama yang membahas konseptualisasi keberlanjutan dalam AI. Baginya, diskusi ini merupakan gelombang ketiga dalam etika AI. Ia menggambarkan gelombang pertama sebagai gelombang yang berfokus pada risiko superintelijen dan pemberontakan robot. Gelombang kedua pada dasarnya adalah apa yang telah kita bahas sejauh ini dalam buku ini, membahas masalah etika seputar pembelajaran mesin dan menghubungkannya dengan praktik, dengan berbagai permasalahan seperti bias, diskriminasi, keterjelasan, dan privasi, di antara yang lainnya. Namun, gelombang ketiga, sebaliknya, "menghadapi bencana lingkungan di zaman kita secara langsung dan secara aktif berupaya melibatkan akademisi, pembuat kebijakan, pengembang AI, dan masyarakat umum dengan dampak lingkungan AI", dengan demikian menempatkan keberlanjutan sebagai elemen inti.

Untuk mengembangkan konseptualisasi keberlanjutannya, Van Wynsberghe membuat perbedaan krusial antara AI untuk keberlanjutan, yaitu menuju tujuan pembangunan berkelanjutan, dan keberlanjutan dalam mengembangkan dan menggunakan sistem AI. Cabang keberlanjutan pertama mungkin yang paling terkenal, karena berupaya menerapkan AI untuk mencapai praktik dan hasil yang berkelanjutan. Misalnya, pembelajaran mesin dapat digunakan untuk mengoptimalkan proses penciptaan energi bersih atau air minum. Yang kedua, dan mungkin yang menimbulkan lebih banyak masalah etika, berkaitan dengan mendorong perubahan dalam keseluruhan siklus pengembangan AI, mulai dari pembangkitan ide hingga tata kelola. Dengan demikian, gagasan keberlanjutan ini bertujuan "menuju integritas ekologis dan keadilan sosial yang lebih besar."

Berdasarkan perbedaan mendasar ini, van Wynsberghe memperkenalkan AI Berkelanjutan melalui penilaian keseluruhan sistem AI sosioteknis. Hal ini tidak hanya melibatkan penerapan teknologi AI, tetapi juga kebutuhan kritis untuk menangani konteks sosioteknis yang lebih luas di mana teknologi ini beroperasi. Definisi yang diusulkan mengarah pada gerakan yang bertujuan untuk mendefinisikan ulang bagaimana teknologi AI

dikembangkan dan digunakan, memastikannya kompatibel dengan sumber daya lingkungan dan nilai-nilai sosial yang berkelanjutan di seluruh generasi sekarang dan mendatang.

AI Berkelanjutan, dari perspektif sosioteknis, mengintegrasikan pertimbangan etika dan keberlanjutan di seluruh ekosistem yang mendukung pengembangan AI. Van Wynsberghe menekankan biaya lingkungan yang substansial terkait dengan AI, khususnya sifat intensif energi dari pelatihan model pembelajaran mendalam. Misalnya, ia merujuk pada sebuah studi oleh Strubell. di mana mereka menunjukkan bahwa pelatihan satu model NLP dapat menghasilkan jejak karbon sekitar 626.155 CO2e (lbs) dibandingkan dengan tingkat konsumsi umum untuk teknologi lain seperti penggunaan mobil dalam 1 masa pakai sebesar 126.000 CO2e (lbs). Ia menunjukkan bahwa jejak karbon dari aktivitas-aktivitas ini signifikan. Emisi dari pelatihan satu model AI sebanding dengan emisi dari lima mobil selama masa pakainya. Perbandingan yang tajam ini menyoroti kebutuhan mendesak bagi komunitas riset AI untuk mempertimbangkan dampak ekologis dari pekerjaan mereka, bukan hanya kemajuan teknologi dan implikasi etis dalam hal diskriminasi, privasi, atau tuntutan moral lainnya.

Lebih lanjut, ia menyerukan peningkatan akuntabilitas dalam industri AI terkait dampak lingkungannya. Seruan Van Wynsberghe untuk bertindak mencakup mengarahkan pendanaan untuk metodologi AI yang lebih berkelanjutan, sehingga memberikan insentif bagi penelitian dan pengembangan yang memprioritaskan pengurangan konsumsi energi dan emisi karbon yang lebih rendah. Pengalihan sumber daya ini dianggap penting untuk mendorong inovasi yang sejalan dengan tujuan mengurangi jejak karbon global dan mencapai pembangunan berkelanjutan. Namun, mencapai hal ini merupakan tugas yang menantang, dan upaya lebih lanjut perlu dilakukan untuk mengintegrasikan keharusan etis ini ke dalam pengambilan keputusan praktis, menyeimbangkan inovasi dan pengembangan teknologi dengan standar keberlanjutan.

Peneliti lain yang telah menyelidiki kekhawatiran ini adalah Crawford dalam bukunya Atlas of AI. Ia menekankan pentingnya konsep-konsep kunci yang dapat mengubah atau memanipulasi persepsi publik tentang AI: “Komputasi tingkat lanjut jarang dipertimbangkan dalam konteks jejak karbon, bahan bakar fosil, dan polusi; metafora seperti “awan” menyiratkan sesuatu yang mengambang dan rapuh di dalam industri alami dan hijau”.

Lebih lanjut, dalam bab “Bumi”, Crawford berpendapat bahwa produksi dan penerapan AI secara intrinsik terkait dengan degradasi ekologi dan penipisan sumber daya yang signifikan. Ia dengan cermat merinci ekstraksi mineral yang diperlukan untuk komponen elektronik yang digunakan untuk membuat perangkat keras yang diperlukan untuk melatih model AI. Kebutuhan ekstraktif ini tidak hanya menyebabkan kerusakan lingkungan, seperti deforestasi dan polusi air, tetapi juga menimbulkan kekhawatiran etis yang mendalam tentang keberlanjutan teknologi AI sejalan dengan kekhawatiran yang dikemukakan oleh van Wynsberghe. Crawford menunjukkan bahwa fakta bahwa sebagian besar sumber daya yang dibutuhkan untuk mendukung sistem AI berasal dari sumber energi tak terbarukan secara langsung bertentangan dengan persepsi efisiensi dan manfaat teknologi tersebut, sehingga mengungkap biaya lingkungan yang tersembunyi akibat pertumbuhan pesat industri ini. Analisisnya juga mencakup limbah elektronik yang dihasilkan dari perangkat keras usang, yang

menumpuk racun di tempat pembuangan sampah (TPA) yang sebagian besar berlokasi di negara-negara kurang berkembang secara ekonomi.

Oleh karena itu, ketidakadilan sosial-ekonomi dan lingkungan yang diabadikan oleh industri AI juga berasal dari praktik-praktik yang tidak berkelanjutan. Crawford berpendapat bahwa beban pengembangan AI, mulai dari eksploitasi tenaga kerja dalam ekstraksi mineral hingga pembuangan limbah elektronik yang sembarangan, secara tidak proporsional memengaruhi negara-negara berkembang. Kesenjangan geografis dalam distribusi biaya lingkungan AI ini menimbulkan isu-isu etika signifikan yang seringkali terlewatkan oleh diskusi "arus utama" tentang gelombang kedua etika AI.

3.2 KRISIS REPLIKASI DALAM AI

Pada akhir tahun 2023, Phillip Ball berkomentar di Nature bahwa penerapan AI yang naif berpotensi berkontribusi pada krisis reproduktifitas di berbagai disiplin ilmu. Dalam artikelnya, ia menyoroti karya berbagai peneliti yang telah menunjukkan bagaimana sistem AI telah gagal. Ball mengklaim hal ini disebabkan, setidaknya sebagian, oleh penggunaan yang tidak tepat dan kesalahpahaman yang meluas tentang AI dan perangkat pembelajaran mesin, terutama dalam fase pelatihan dan pengujiannya. Diskusi ini menyoroti dua aspek krusial: penggunaan teknologi AI yang inovatif dalam konteks baru dan potensi jebakan penerapan AI tanpa validasi yang ketat.

Contoh jebakan ini adalah kasus sistem pembelajaran mesin yang digunakan untuk menganalisis citra sinar-X untuk mendeteksi COVID-19 selama pandemi. Seiring berlanjutnya pandemi, alat tes menjadi langka; oleh karena itu, para peneliti di India mengusulkan agar mereka dapat melatih model pembelajaran mesin untuk belajar dari pemindaian sinar-X dada guna membedakan antara pasien yang terinfeksi dan yang tidak terinfeksi. Makalah tersebut dikutip beberapa kali; namun, hampir setahun setelah publikasinya, Dhar dan Shamir melatih algoritma serupa menggunakan sebagian kecil citra yang sama dengan latar belakang kosong saja, yaitu, tidak menunjukkan bagian tubuh apa pun. Anehnya, algoritma yang dilatih ulang tersebut juga mampu "mendeteksi" kasus COVID-19.

Masalahnya tampaknya terletak pada adanya perbedaan yang konsisten pada latar belakang citra medis dalam kumpulan data. Sistem AI dapat menangkap artefak tersebut untuk berhasil dalam tugas diagnostik tanpa mempelajari fitur klinis apa pun yang relevan, sehingga membuatnya tidak berguna secara medis. Investigasi mereka mengungkapkan bahwa AI dapat mengidentifikasi COVID-19 dari bagian-bagian gambar sinar-X yang tidak mengandung informasi diagnostik, melainkan hanya noise latar belakang. Penemuan ini menunjukkan masalah mendasar dalam model pembelajaran mesin: risiko menghasilkan hasil yang menyesatkan akibat bias atau anomali yang tidak dikenali dalam data pelatihan.

Masalah lain yang ditunjukkan dalam analisis Ball adalah kasus kebocoran data, yang terjadi ketika informasi yang asing secara keliru dimasukkan selama proses pelatihan model. Fenomena ini dapat terjadi ketika data yang ditujukan hanya untuk set uji, yang berfungsi untuk mengevaluasi model, secara tidak sengaja dimasukkan ke dalam set data pelatihan. Kebocoran data juga dapat terjadi ketika langkah-langkah prapemrosesan, seperti duplikasi

baris untuk oversampling, diterapkan ke seluruh set data, alih-alih terbatas pada subset pelatihan. Akibatnya, kebocoran data mengakibatkan estimasi kinerja model yang terlalu tinggi, karena model tersebut tampak unggul pada data yang seharusnya tidak terpapar selama fase pelatihan. Kapoor dan Narayanan menyoroti isu-isu signifikan terkait kebocoran data dalam penelitian pembelajaran mesin di 17 disiplin ilmu. Dalam studi mereka, mereka mengungkapkan bahwa kebocoran data menyebabkan penilaian kinerja model yang terlalu optimistis dan hasil yang tidak dapat direproduksi, yang berkontribusi pada krisis reproduktifitas.

Para penulis memberikan taksonomi kebocoran data yang terperinci, membedakan delapan jenis spesifik, mulai dari kesalahan sederhana dalam penanganan data hingga isu-isu kompleks yang memerlukan penelitian lebih lanjut untuk sepenuhnya memahami dan memitigasinya. Untuk mengatasi isu-isu ini, Kapoor dan Narayanan memperkenalkan lembar informasi model yang dirancang untuk membantu para peneliti mendokumentasikan dan memverifikasi secara sistematis bahwa model mereka bebas dari kebocoran data. Dengan demikian, para penulis berpendapat, memvalidasi model AI terhadap standar yang kuat dapat berkontribusi positif bagi masyarakat, tetapi penerimaan yang tidak kritis terhadap hasil yang dihasilkan AI tanpa pengawasan menyeluruh, terutama di area dengan dampak sosial yang tinggi, mengkhawatirkan.

Melengkapi kritik sebelumnya, Bausell mengartikulasikan bahwa krisis reproduktifitas dalam penelitian ilmiah tidak boleh dianggap hanya sebagai insiden yang terisolasi tetapi sebagai manifestasi dari kegagalan sistemik. Ia berpendapat bahwa kegagalan-kegagalan ini berakar pada metodologi penelitian yang cacat dan budaya yang berlaku yang secara tidak proporsional lebih mementingkan hasil sensasional daripada investigasi yang cermat dan teliti. Aspek ini khususnya relevan dalam konteks AI, di mana daya tarik dan kebaruan penggunaan metode pembelajaran mesin seringkali mengalahkan ketelitian metodologis yang ketat.

Lebih lanjut, Bausell membahas bagaimana bias publikasi, ketika jurnal menunjukkan preferensi untuk menerbitkan hasil positif daripada hasil negatif atau nihil, berkontribusi pada catatan ilmiah yang terdistorsi. Bias ini semakin nyata dalam penelitian yang berkaitan dengan atau yang memanfaatkan AI, melanggengkan siklus sensasi seputar teknologi yang sedang berkembang. Demikian pula, Bausell menyoroti praktik-praktik penelitian yang dipertanyakan, seperti pelaporan selektif dan penghentian opsional, yang ia akui bukan sekadar kesalahan penilaian, tetapi seringkali merupakan produk dari tekanan dan insentif institusional. Praktik-praktik ini mengakar kuat di beberapa lingkungan penelitian, yang menyebabkan penerimaan atau pengawasan pasif. Skenario ini mendorong kurangnya ketelitian dalam penelitian AI dan integrasinya di berbagai domain ilmiah, sehingga memperburuk tantangan yang ada dalam reproduktifitas ilmiah. Kritik Bausell menggarisbawahi perlunya pergeseran budaya dan metodologi dalam komunitas ilmiah untuk mengatasi masalah sistemik ini secara efektif.

Menilai krisis replikasi dalam sains, secara lebih umum, juga memiliki dimensi filosofis yang berkaitan dengan standar etika yang perlu kita tegakkan untuk menghindari krisis yang disebutkan. Romero menganalisis isu ini dari perspektif Filsafat Sains. Ia membahas tiga solusi kritis untuk krisis replikasi, mengklasifikasikannya menjadi reformasi statistik, reformasi

metodologi, dan reformasi sosial. Setiap kategori membahas aspek-aspek berbeda dari kelemahan sistemik yang berkontribusi terhadap krisis, menekankan integrasi etika penelitian dan pandangan epistemologi sosial untuk mengatasi krisis replikasi.

1. **Reformasi Statistik:** Reformasi ini menyerukan perubahan mendasar dalam cara data dianalisis dalam penelitian ilmiah. Proposal yang diajukan mencakup adopsi statistik Bayesian, yang menuntut asumsi transparan dan menawarkan inferensi langsung dari hipotesis nol, sehingga membantu mengatasi kegagalan replikasi. Proposal penting lainnya adalah menurunkan ambang batas nilai-p dari 0,05 menjadi 0,005, yang bertujuan untuk mengurangi positif palsu dan meningkatkan ketelitian statistik dari studi yang dipublikasikan. Selain itu, penekanan pada ukuran efek dan interval kepercayaan di atas signifikansi statistik semata dapat mengalihkan fokus ke implikasi praktis dari temuan penelitian, sehingga meningkatkan penerapan dan reliabilitasnya.
2. **Reformasi Metodologis**: Reformasi ini menargetkan prosedur pelaksanaan dan pelaporan penelitian. Pra-registrasi penelitian dipromosikan untuk mengekang pelaporan selektif dan p-hacking dengan berkomitmen pada rencana awal peneliti sebelum pengumpulan data dimulai. Penerapan praktik sains terbuka, termasuk berbagi data dan materi, bertujuan untuk meningkatkan transparansi dan memfasilitasi verifikasi serta replikasi karya ilmiah. Laporan terdaftar, sebuah format publikasi baru, melibatkan tinjauan sejawat sebelum hasilnya diketahui; pendekatan ini mengurangi bias publikasi karena penerimaan didasarkan pada pertanyaan penelitian dan ketelitian metodologis, bukan pada kebaruan atau signifikansi hasil.
3. **Reformasi Sosial:** Dengan mengatasi faktor budaya dan kelembagaan yang menghambat replikasi, reformasi ini menyarankan perombakan sistem penghargaan akademik untuk mengakui dan memberi insentif pada upaya replikasi. Mengalokasikan dana khusus untuk studi replikasi dan menyesuaikan kriteria jabatan dan promosi akademik untuk menghargai replikasi dan ketelitian metodologis atas publikasi temuan baru merupakan strategi yang bertujuan untuk menyelaraskan kembali insentif ilmiah dengan praktik yang mendorong luaran penelitian yang kuat dan andal.

Dengan cara yang lebih optimis, tetap pada argumen yang sama, berpendapat bahwa perdebatan reproduktifitas harus dilihat sebagai peluang, bukan krisis, karena hal ini menyoroti area dalam kerangka kerja penelitian ilmiah yang dapat ditingkatkan untuk ketahanan dan integritas di masa mendatang. Perspektif ini berasal dari potensi untuk menerapkan perubahan sistemik yang meningkatkan kualitas dan kredibilitas luaran penelitian. Pendekatan proaktif dan berwawasan ke depan ini menyoroti pengembangan yang konstruktif, alih-alih berkutat pada kekurangan praktik saat ini.

Lebih lanjut, dalam sebuah laporan oleh OCDE, terdapat bagian yang didedikasikan untuk peningkatan reproduktifitas dalam penelitian AI, yang bertujuan untuk meningkatkan kepercayaan dan produktivitas. Di sana, Gundersen menunjukkan bahwa hampir 70% penelitian AI mungkin tidak dapat direproduksi, menekankan bahwa sumber-sumber ketidakreproduksian tersebut mencakup desain studi, pilihan algoritma pembelajaran mesin, proses penanganan data, serta evaluasi dan pelaporan temuan penelitian. Faktor-faktor ini

dapat menyebabkan disparitas yang signifikan dalam replikasi di berbagai lingkungan atau ketika metodologi yang berbeda diterapkan, yang karenanya Gundersen mengusulkan serangkaian rekomendasi untuk meningkatkan reproduktifitas penelitian AI.

Penulis menyarankan agar lembaga penelitian menerapkan praktik penelitian AI terbaik, termasuk pelatihan menyeluruh dan proses jaminan kualitas. Penerbit didorong untuk menstandardisasi proses peninjauan mereka dan mewajibkan publikasi kode dan data di samping temuan penelitian untuk memfasilitasi verifikasi. Lembaga pendanaan disarankan untuk memprioritaskan transparansi dan penelitian terbuka dengan mewajibkan penelitian yang mereka danai dipublikasikan dalam format akses terbuka dan semua keluaran penelitian, termasuk kode dan data, dibagikan secara bebas.

Sebagaimana dijelaskan oleh beberapa akademisi, wacana seputar krisis reproduktifitas dalam penelitian AI tidak hanya mengungkap tantangan inheren tetapi juga menekankan kebutuhan mendesak akan pertimbangan etis. Penerapan AI yang naif di bidang-bidang seperti diagnostik medis tanpa validasi yang ketat menghasilkan hasil yang menyesatkan. Kasus ini menggarisbawahi implikasi etis kritis dari ketergantungan pada sistem AI yang mungkin mendasarkan hasil diagnostik pada fitur data yang tidak relevan, sehingga berpotensi membahayakan nyawa akibat penilaian medis yang salah. Insiden semacam itu menyoroti tanggung jawab yang mendalam untuk memastikan bahwa sistem AI tidak hanya mahir secara teknis tetapi juga dikembangkan secara etis untuk mencegah kerugian dan meningkatkan kepercayaan.

Dalam mengatasi masalah ini, komunitas ilmiah yang lebih luas, termasuk entitas seperti Jaringan Reproduktifitas Inggris dan OECD, mengadvokasi reformasi multidimensi yang mencakup domain statistik, metodologis, dan sosial. Reformasi ini bertujuan untuk mengkalibrasi ulang paradigma ilmiah dengan memprioritaskan standar etika dan praktik penelitian yang kuat di atas sensasionalisme dan kebaruan. Penekanan pada penerapan AI yang etis, proses validasi yang ketat, dan transparansi praktik penelitian tidak hanya meningkatkan reproduktifitas dan keandalan keluaran ilmiah, tetapi juga memastikan bahwa kemajuan AI berkontribusi positif terhadap kesejahteraan masyarakat. Pendekatan holistik untuk mengatasi krisis reproduktifitas dalam penelitian AI ini menggarisbawahi peran penting etika dalam memandu kemajuan teknologi, memastikan bahwa AI berfungsi sebagai alat penelitian ilmiah yang bermanfaat.

BAB 4
BIAS DALAM AI

4.1 PENDAHULUAN

Munculnya AI dan beragam aplikasinya dalam kehidupan sehari-hari telah mendorong upaya untuk lebih memahami potensi risiko yang terkait dengan penggunaannya. Pengaruhnya terhadap kehidupan kita sehari-hari diperkirakan akan meningkat secara signifikan. AI membantu kita setiap hari dalam pengambilan keputusan, mulai dari memilih film dan menerjemahkan teks hingga merekrut staf untuk sebuah perusahaan. Penggunaannya memiliki manfaat yang jelas, terutama dipicu oleh kemampuan pembelajaran induktif untuk menganalisis data dalam jumlah besar dan mengekstrak pola statistik halus yang mustahil dideteksi hanya dengan upaya manusia. AI menghemat waktu kita dengan memproses data dalam jumlah besar secara efisien dan mengidentifikasi informasi relevan yang bermanfaat bagi kita.

Berbagai keunggulan AI sangat bergantung pada algoritma pembelajaran mesin yang diandalkannya. Pembelajaran mesin adalah subbidang AI yang berfokus pada pengembangan algoritma pembelajaran induktif. Algoritma ini menggunakan data untuk mengidentifikasi pola, yang dikodekan oleh algoritma menjadi model. Misalnya, pengklasifikasi menghasilkan model untuk menyimpulkan variabel kategoris target, yang dikenal sebagai variabel target, untuk berbagai jenis objek. Misalnya, sistem pengenalan angka (Pengenalan Karakter Optik) mengandalkan pengklasifikasian citra angka-angka tersebut untuk mempelajari keteraturan data. Keteraturan ini berkorelasi dengan variabel target, dan pengklasifikasi dapat menggeneralisasi ke instans baru dari asosiasi ini. Model dikatakan belajar karena mengembangkan kapasitas untuk menyimpulkan kelas objek yang tidak digunakan selama pelatihan pengklasifikasi.

Karena pembelajaran mesin didasarkan pada pembelajaran asosiasi antar deskriptor objek yang digunakannya, penting untuk menanyakan apa yang dikodekan oleh asosiasi ini. Risikonya adalah asosiasi ini dapat memperkuat bias yang memengaruhi individu, kelompok, atau subkelompok dalam masyarakat, dan oleh karena itu, keputusan yang didasarkan pada salah satu sistem ini pada akhirnya bisa menjadi tidak adil. Meskipun terdapat beragam definisi keadilan dalam AI, dan banyak di antaranya berasal dari tradisi yang kuat di bidang lain seperti filsafat dan ilmu sosial, esensi dasar dari semua definisi ini adalah tidak adanya prasangka atau favoritisme yang tidak beralasan terhadap individu atau kelompok berdasarkan karakteristik inheren mereka.

Oleh karena itu, model pembelajaran mesin yang tidak adil dapat disebabkan oleh ketergantungan pada keteraturan yang bias terhadap kelompok atau individu tertentu. Kami akan mengilustrasikan situasi ini dengan dua contoh. Pertama, sistem AI yang dirancang untuk sistem peradilan AS, yang telah kami perkenalkan di Bab 1, diikuti oleh kontes kecantikan di mana AI menjadi hakim resmi.

Meninjau Kembali Perangkat Lunak COMPAS

Compas (Compas) Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) adalah perangkat lunak manajemen kasus yang dikembangkan oleh Northpointe, Inc. (sekarang Equivant). Pengadilan di New York, Wisconsin, California, dan Florida menggunakannya. Perangkat lunak ini dirancang untuk menilai kemungkinan seseorang mengulangi tindak pidana. COMPAS menetapkan skor risiko berdasarkan deskriptor perilaku manusia.

Pengembangannya dimotivasi oleh keinginan untuk meminimalkan bias kognitif dan pengaruh prasangka dalam pengambilan keputusan pengadilan, salah satu yang paling menonjol adalah "efek hakim lapar". Efek ini menyoroti bagaimana hakim cenderung lebih lunak setelah makan siang dan lebih ketat sebelum makan siang. Temuan terkait menunjukkan bahwa kasus yang lebih kompleks dijadwalkan untuk pagi hari, sementara kasus yang lebih sederhana ditangani pada sore hari karena durasi yang diperkirakan.

Pada bulan Juli 2016, Mahkamah Agung Wisconsin memutuskan bahwa skor risiko COMPAS dapat dipertimbangkan selama penjatuhan hukuman. Namun, kritik segera muncul. Kekhawatiran utamanya adalah, sebagai perangkat lunak berpemilik, COMPAS bersifat algoritmik yang tidak transparan dan tidak dapat diaudit atau diperiksa secara publik. Asimetri informasi ini merugikan individu yang dianalisis, karena mereka tidak memiliki akses ke data dasar yang mendukung kalimat mereka, sehingga mempersulit kemampuan mereka untuk mengajukan banding dan hak mereka atas penjelasan (sesuatu yang sekarang eksplisit dalam Undang-Undang AI Uni Eropa).1 Fondasi kalimat semacam itu dapat diatasi menggunakan model "kotak putih", sebuah sistem terbuka yang dapat diperiksa. Namun, baik sistem transparan maupun sistem terbuka berbasis pembelajaran mesin menghadapi tantangan karena algoritmanya bergantung pada data.

Investigasi ProPublica terhadap COMPAS menemukan bias terhadap terdakwa kulit hitam. Studi tersebut mengungkapkan bahwa orang kulit berwarna hampir dua kali lebih mungkin dilabeli berisiko tinggi oleh COMPAS dibandingkan orang kulit putih, namun banyak dari individu yang dianggap berisiko tinggi ini tidak melakukan kejahatan kekerasan lagi. Sebaliknya, terdakwa kulit putih hampir dua kali lebih mungkin dilabeli berisiko rendah dibandingkan orang kulit hitam, namun sejumlah besar kasus ini kembali melakukan kejahatan. Tren yang diprediksi, yang diidentifikasi oleh algoritma dari data historis, dapat dikaitkan dengan pola diskriminasi yang terdokumentasi dengan baik terhadap orang kulit hitam dalam pengawasan, penahanan, dan pemenjaraan di Amerika Serikat, tempat data dikumpulkan.

Kontes kecantikan

Contoh lain bias dalam AI ditemukan di Beauty.AI, yang memperkenalkan Kontes Kecantikan Internasional Pertama yang dinilai oleh AI. Peserta mengunduh aplikasi untuk mengirimkan swafoto, yang dievaluasi oleh algoritma untuk menentukan pemenang, yang akan dipromosikan di berbagai media berita global.

Setelah pengumuman pemenang pada tahun 2016, kontroversi pun muncul. Hasilnya menyoroti masalah yang signifikan: selain menggunakan faktor-faktor seperti proporsi wajah,

simetri, dan kerutan untuk menilai daya tarik, AI tampaknya kurang disukai individu berkulit gelap. Di antara 6.000 peserta dari lebih dari 100 negara, termasuk India dan Afrika, 40 posisi teratas didominasi oleh orang Kaukasia, dengan hanya satu kontestan 40 teratas yang berkulit gelap.

Meskipun bukti dari kontes kecantikan tersebut mungkin dianggap anekdotal, kontroversi tersebut menggarisbawahi dampak kritis bias dalam keluaran algoritmik. Bergantung pada konteks penggunaannya, konsekuensi dari bias ini dapat sangat merugikan.

Kebutuhan untuk mengatasi bias dalam AI

Ketidakadilan algoritmik dapat muncul dari berbagai sumber, terutama bias yang berasal dari data dan bias yang berasal dari algoritma itu sendiri. Terlepas dari manfaat AI, penerapan sistem ini memerlukan tanggung jawab yang signifikan. Tanggung jawab ini mencakup penanggulangan dampak bias algoritmik dan penanganan yang tepat terhadap dampak buruk yang mungkin ditimbulkannya di masyarakat. Dampak bias algoritmik dapat dikurangi dengan identifikasi awal sumbernya, sehingga memungkinkan pengurangan dampaknya.

4.2 BIAS DALAM AI

Mendefinisikan bias bukanlah tugas yang mudah. Dalam konteks statistik, bias bisa berarti perilaku sistem AI yang tidak akurat. Dalam konteks hukum, bias berfokus pada aspek-aspek dampak yang berbeda. Dalam konteks kognitif dan sosial, pengambilan keputusan manusia merupakan pengaruh yang lebih relevan. Di sini, kami tidak terlalu spesifik tentang konteks interpretasi bias, melainkan mempertimbangkan pengaruh kontekstual terkait beberapa definisi bias yang telah mapan. Kami mengenali setidaknya tiga sifat yang mencirikan bias dalam AI berdasarkan karya Zhai dan Krajcik.

Pertama, bias menyiratkan kesalahan, dengan kata lain, penyimpangan antara observasi dan kebenaran dasar. Bias juga membutuhkan komponen sistematis; bias merupakan kesalahan sistematis dan bukan sekadar kejadian acak. Lebih lanjut, ketika kita merujuk pada bias, kita juga menyinggung kecenderungan yang tidak beralasan atau tidak relevan yang mendukung atau menentang beberapa ide atau entitas dibandingkan yang lain. Kini, dengan sepakat bahwa bias merupakan kesalahan sistematis dengan kecenderungan prasangka yang mendasarinya, asal-usulnya dapat beragam, dan kita mengenali setidaknya tiga konteks asal: sosial, teknis, dan kognitif.

Bias Sosial, Teknis, dan Kognitif

Bias sosial dalam AI berasal dari konteks sosial, budaya, dan kelembagaan tempat data dihasilkan dan digunakan. Bias ini mencerminkan dan melanggengkan prasangka dan ketidaksetaraan yang ada dalam masyarakat. Sistem AI sering kali belajar dari kumpulan data besar yang dihasilkan oleh manusia. Jika kumpulan data ini mengandung bias sosial, seperti disparitas ras, gender, atau ekonomi, sistem AI kemungkinan besar akan menangkap dan mereplikasi bias ini. Bias yang tertanam dalam data ini mencerminkan prasangka historis dan budaya, yang mengarah pada hasil diskriminatif ketika sistem AI diterapkan. Misalnya, sistem AI yang digunakan dalam perekrutan mungkin lebih mengutamakan kelompok demografis

tertentu daripada yang lain jika dilatih dengan data yang mencerminkan praktik perekrutan yang bias. Oleh karena itu, kami menyebut bias sosial sebagai pengaruh historis, institusional, dan sosial yang sistematis, yang mencerminkan perilaku diskriminatif, asumsi, dan ketimpangan struktural yang ada.

Bias teknis muncul dari metode, proses, dan perangkat spesifik yang digunakan untuk mengembangkan dan menerapkan sistem AI. Bias ini terkait dengan kendala algoritmik dan keputusan teknis pengembang, seperti pemilihan data dan desain model. Ketika data yang digunakan untuk melatih model AI tidak mewakili keseluruhan populasi, bias dapat terjadi. Misalnya, penggunaan data dari satu sumber yang homogen dapat menimbulkan bias yang memengaruhi kinerja AI di berbagai skenario. Desain teknis sistem AI, termasuk pemilihan fitur, arsitektur model, dan kriteria optimasi, dapat menimbulkan bias jika tidak dipertimbangkan dengan cermat. Bias teknis juga dapat diakibatkan oleh kesalahan atau keterbatasan dalam algoritma, seperti overfitting pola tertentu dalam data pelatihan atau kegagalan untuk menggeneralisasi ke konteks baru. Oleh karena itu, yang kami maksud dengan bias teknis adalah kesalahan dalam representasi atau kesalahan statistik sistematis yang melibatkan tingkat keberpihakan atau konsekuensi diskriminatif tertentu.

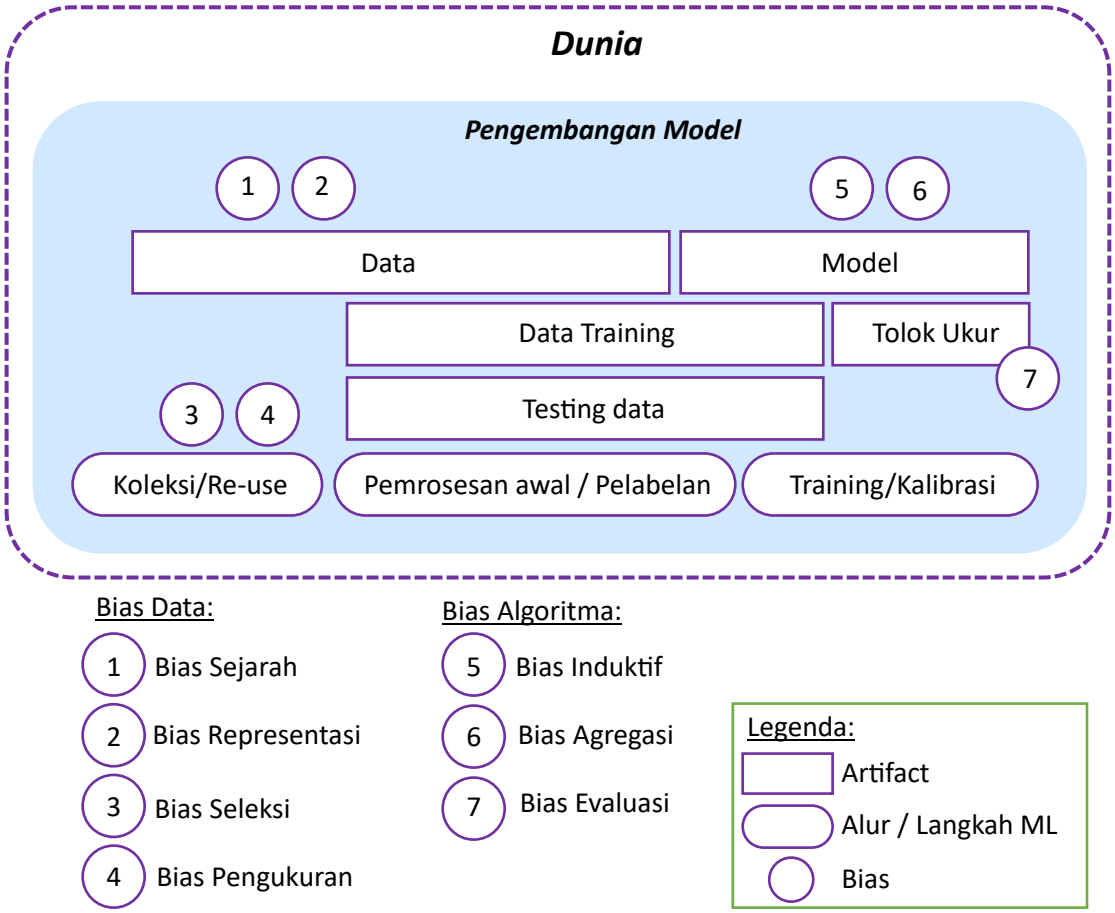
Terakhir, bias kognitif merujuk pada kesalahan manusia dalam penilaian atau penalaran yang dapat memengaruhi fungsi dan penerapan sistem AI. Bias ini terjadi ketika sistem AI mencerminkan keterbatasan kognitif dan heuristik pengembang dan penggunanya, ketika sistem tersebut meniru pola pengambilan keputusan manusia, atau ketika manusia membuat keputusan metodologis yang bias yang berkontribusi pada keberadaan bias teknis, seperti bias seleksi (lihat bagian 4.2.2).

Sistem AI sering kali belajar dari interaksi dengan manusia, yang dapat menimbulkan bias kognitif jika manusia yang terlibat dalam pengembangan atau penggunaan sistem tersebut memiliki keyakinan yang keliru atau membuat penilaian yang keliru. Data pelatihan yang mencakup pola pengambilan keputusan manusia dapat menanamkan bias kognitif dalam sistem AI. Jika manusia secara konsisten membuat keputusan yang bias dalam konteks tertentu, AI yang dilatih dengan data ini dapat mereplikasi bias tersebut. Misalnya, AI dapat belajar untuk "mengobjektifikasi" tubuh perempuan jika orang yang memberi label pada gambar tubuh perempuan menganggapnya lebih sugestif secara seksual daripada gambar tubuh laki-laki.

Contoh lain dapat ditemukan dalam bias data medis yang umum diketahui, yang berasal dari bias profesional dalam spesialisasi medis tertentu. Bias profesional dalam sistem AI dapat muncul ketika sistem dilatih dengan data yang mencerminkan bias kognitif para profesional yang terlibat dalam pengembangan atau penggunaannya. Bias ini dapat secara tidak sengaja tertanam dalam model AI ketika pola pengambilan keputusan manusia dimasukkan ke dalam data pelatihan. Misalnya, jika tenaga kesehatan profesional memiliki bias yang konsisten dalam mendiagnosis kondisi tertentu lebih sering pada kelompok demografi tertentu, sistem AI yang dilatih dengan data tersebut dapat mereplikasi dan bahkan memperkuat bias ini. Misalnya, bias rasial berkaitan dengan praktisi medis yang mengabaikan fitur diagnostik, yang menyebabkan kelompok non-kulit putih tertentu kurang terdiagnosis

atau salah didiagnosis dibandingkan dengan pasien kulit putih, serta memiliki tingkat akses yang berbeda terhadap layanan kesehatan.

Hal ini mungkin disebabkan oleh bias yang menyebabkan penyedia layanan kesehatan meremehkan atau mengabaikan gejala yang dilaporkan oleh kelompok-kelompok ini, serta bias historis terhadap nyeri yang dialami perempuan. Perempuan, terutama dalam konteks ginekologi, sering melaporkan bahwa nyeri mereka diabaikan atau diremehkan oleh penyedia layanan kesehatan. Kondisi seperti endometriosis, sindrom ovarium polikistik, atau nyeri panggul kronis seringkali kurang terdiagnosis atau salah didiagnosis karena keluhan nyeri yang dialami wanita tidak ditanggapi seserius yang seharusnya.



Gambar 4.1: Contoh bias yang dapat memengaruhi model pembelajaran mesin.

Dengan demikian, sistem AI dapat mencerminkan bias kognitif, terutama jika pengembang dan orang yang menghasilkan data tanpa sadar mengodekan bias ini ke dalam sistem. Oleh karena itu, dengan bias kognitif, kami secara luas mengacu pada kesalahan manusia sistematis yang terkait dengan bias implisit dan heuristik serta pengaruh keputusan manusia terhadap proses pengembangan AI.

Dengan adanya perbedaan ini, kini kita dapat mengkaji lebih dekat beberapa bias paling umum dalam pembelajaran mesin. Bias-bias tersebut mencakup bias yang muncul dari pengumpulan data (bias data) dan bias yang muncul dan diperburuk selama fase pelatihan (bias algoritmik). Gambar 4.1 menunjukkan proses pembelajaran mesin yang disederhanakan

beserta masukan (data) dan keluaran (model) utamanya. Kami menggunakan Gambar ini untuk mengilustrasikan di mana bias muncul dalam pembelajaran mesin.

Bias Data

Sebelum mengeksplorasi jenis-jenis bias data secara spesifik, penting untuk memahami bahwa bias dalam data dapat berasal dari salah satu dari tiga sumber utama: sosial, teknis, dan kognitif. Bias-bias ini dapat secara signifikan memengaruhi hasil model pembelajaran mesin, yang mengarah pada hasil yang bias atau tidak adil. Pengumpulan data, yang mendahului analisis data eksploratif dan pelatihan model, merupakan fase kritis di mana bias semacam itu dapat muncul. Selama fase ini, populasi target diidentifikasi, dan biasanya, sampel diambil dari populasi ini karena ketidakpraktisan dan biaya yang terkait dengan pengumpulan data tingkat sensus.

Terdapat berbagai jenis bias data, seperti yang ditunjukkan pada Gambar 4.1. Kita akan membahasnya untuk memahami sifatnya.

Bias Historis

Bias historis dalam data mencerminkan pola ketimpangan struktural dan diskriminasi yang mengakar dan telah berkembang seiring waktu. Bias ini dapat bermanifestasi dalam berbagai bentuk, seperti peningkatan pengawasan di lingkungan yang didominasi penduduk kulit hitam di AS, standar kecantikan yang lebih mengutamakan karakteristik Kaukasia daripada ras lain, atau kesenjangan upah gender yang terus-menerus di pasar tenaga kerja.

Bias historis muncul dari pola ketidaksetaraan, diskriminasi, atau stereotip yang mengakar. Pola-pola ini telah mengakar dalam masyarakat dari waktu ke waktu dan menemukan jalannya ke dalam sistem AI, terutama melalui data pelatihan. Bias ini tertanam dalam data yang dikumpulkan dari konteks historis, yang mencerminkan norma sosial, praktik, dan dinamika kekuasaan yang ada saat data dikumpulkan. Tidak seperti bentuk bias lain yang mungkin diakibatkan oleh desain eksperimen yang cacat atau kesalahan pengambilan sampel, bias historis terjadi bahkan ketika proses pengumpulan data secara metodologis baik. Oleh karena itu, masalah utama dengan bias historis adalah bahwa ia mencerminkan dunia sebagaimana adanya, alih-alih "bagaimana seharusnya". Hal ini dapat melanggengkan ketidakadilan dan ketidaksetaraan masa lalu ketika data yang bias ini digunakan untuk melatih model pembelajaran mesin yang digunakan untuk membuat atau menginformasikan keputusan tentang masa depan.

Proses pengumpulan data merupakan tahap kritis di mana bias semacam itu dapat muncul. Misalnya, dalam sistem NLP, representasi kata dibangun melalui pra-pelatihan pada teks dalam jumlah besar, sehingga menciptakan penyisipan kata. Teks-teks ini seringkali bersumber dari repositori publik seperti Wikipedia, Google News, dan Book Corpus. Karena sumber-sumber ini mencerminkan konteks sosial pada saat penciptaannya, sumber-sumber ini mau tidak mau mereproduksi stereotip dan bias yang lazim pada era tersebut.

Penelitian telah menunjukkan bahwa stereotip gender, misalnya, dapat menyebabkan asosiasi yang bermasalah seperti "laki-laki untuk dokter sebagaimana perempuan untuk perawat". Meskipun hal ini mungkin tampak anekdotal, implikasinya sangat luas, karena sistem yang didasarkan pada representasi bias ini dapat menyebabkan kerugian yang

signifikan. Sebagaimana disorot dalam karya Noble, mesin pencari telah terbukti mengabadikan representasi stereotip individu kulit hitam. Hal ini terjadi karena representasi teks dalam sistem NLP pada dasarnya bergantung pada hubungan spasial antara penyisipan kata dalam ruang laten.

Oleh karena itu, analogi kata menguraikan hubungan yang telah dipelajari mesin dari data bias dan menggunakan hubungan ini untuk menghasilkan hasil. Saat ini, aplikasi berbasis NLP terintegrasi secara luas ke dalam berbagai sistem yang berinteraksi dengan manusia sehari-hari, seperti chatbot, sistem penerjemahan otomatis, pengenalan suara, dan mesin pencari web. Akibatnya, representasi yang bias ini telah memengaruhi cara pengguna menerima dan menginterpretasikan hasil dari berbagai layanan yang dimediasi AI, menggarisbawahi pentingnya mengatasi bias historis dalam data.

Bias Representasi

Bias representasi terjadi ketika kumpulan data yang digunakan untuk melatih model pembelajaran mesin gagal menangkap keragaman dan variabilitas populasi target secara akurat, sehingga mengakibatkan generalisasi yang buruk untuk subset tertentu dari populasi. Bias ini dapat muncul dalam beberapa cara dan dapat mengakibatkan hasil yang miring atau tidak merata, di mana model berkinerja baik untuk kelompok yang terwakili secara berlebihan. Namun, model berkinerja buruk atau membuat prediksi yang salah untuk kelompok yang kurang terwakili.

Sebagaimana disoroti oleh Suresh dan Guttag, bias representasi dapat muncul selama pendefinisian populasi target jika tidak mencerminkan populasi tujuan penggunaan secara memadai. Misalnya, data yang mewakili Brasil mungkin tidak dapat digeneralisasi ke populasi Jepang, atau data dari Hanover, Jerman, 30 tahun yang lalu, mungkin tidak secara akurat mewakili populasi saat ini. Demikian pula, jika populasi target mengandung kelompok-kelompok tertentu yang secara alami kurang terwakili, seperti individu hamil dalam kumpulan data medis orang dewasa, model tersebut mungkin kurang robust untuk kelompok minoritas ini karena terbatasnya data yang tersedia tentang mereka.

Lebih lanjut, bias representasi dapat terjadi selama proses pengambilan sampel jika metode yang digunakan terbatas atau tidak merata, sehingga menghasilkan sampel pengembangan yang merepresentasikan subset populasi target yang miring [284]. Misalnya, dalam pemodelan penyakit menular, jika data medis hanya tersedia untuk individu yang dianggap cukup serius untuk skrining lebih lanjut, sampel yang dihasilkan akan bias terhadap kasus yang lebih parah, sehingga menghasilkan model yang mungkin tidak berkinerja baik di seluruh populasi.

Dalam beberapa kasus, data mungkin secara akurat merepresentasikan populasi target namun tetap menunjukkan bias representasi. Hal ini terjadi karena jika variabel deskriptif kelompok relevan dengan tugas, algoritma pembelajaran mesin memerlukan sampel yang seimbang terkait deskriptor ini. Ketidakseimbangan kelas kemudian dikenali sebagai masalah dalam pelatihan pengklasifikasi, karena sampel yang tidak seimbang berdasarkan variabel target meningkatkan risiko overfitting terhadap kelas mayoritas.

Lebih lanjut, bias representasi dapat muncul karena efek transfer learning. Jika model yang dilatih pada populasi target kemudian diterapkan pada populasi yang berbeda, model tersebut akan mereplikasi stereotip populasi target ke populasi yang baru. Misalnya, dalam masalah deteksi bot, sistem yang banyak digunakan di Twitter (sekarang X) seperti Botometer, yang dilatih pada data dari populasi Anglo-Saxon, menghasilkan hasil yang tidak konsisten ketika diterapkan pada populasi Hispanik.

Sistem ini cenderung merepresentasikan volume bot yang sebenarnya di negara-negara berbahasa Spanyol secara berlebihan. Perbedaan ini terjadi karena pemindahan model dari satu konteks linguistik dan budaya ke konteks lain mengabaikan perbedaan budaya, penggunaan idiomatis lokal, dan pola perilaku spesifik dari konteks baru yang tidak diamati dalam populasi target. Dalam hal ini, bias representasi disebabkan oleh kurangnya representasi perbedaan budaya dan linguistik yang unik dari konteks baru tersebut.

Oleh karena itu, bias representasi muncul dari kekurangan metodologis selama proses pengumpulan sampel. Jenis bias ini dikaitkan dengan penggambaran yang tidak adil atau tidak memadai terhadap strata dan kelompok masyarakat tertentu.

Bias seleksi

Bias seleksi mengacu pada kejadian di mana data yang digunakan untuk melatih sistem AI tidak mewakili realitas yang ingin dimodelkannya. Dengan kata lain, ketika data pelatihan model pembelajaran mesin tidak mewakili lingkungan tempat model beroperasi karena cara data dipilih atau dikumpulkan.

Salah satu jenis bias seleksi yang umum adalah seleksi mandiri. Hal ini sering muncul dalam sistem rekomendasi, di mana pengguna memilih sendiri untuk mengekspresikan preferensi mereka terhadap item tertentu yang mereka pilih untuk dinilai. Karena tujuan utama sistem rekomendasi adalah untuk menyarankan konten yang relevan kepada pengguna di suatu platform, sistem ini sangat bergantung pada log konten yang sebelumnya telah disukai pengguna. Strategi penyaringan kolaboratif, yang merupakan bagian integral dari sistem ini, menghasilkan rekomendasi berdasarkan preferensi bersama di antara pengguna. Prinsip dasarnya adalah jika sekelompok pengguna memiliki selera yang sama, item yang disukai oleh beberapa anggota kelompok tetapi belum ditemukan oleh yang lain dapat direkomendasikan kepada yang terakhir. Akibatnya, pengguna lebih mungkin terpapar, dan berpotensi menilai item yang telah disukai oleh orang lain. Hal ini juga menciptakan lingkaran umpan balik, di mana hanya sebagian kecil item yang mengumpulkan data preferensi, sementara item lain tetap tidak diberi peringkat, sebagian karena item tersebut tidak pernah ditampilkan kepada sebagian besar pengguna.

Serupa dengan itu, platform konten, seperti layanan streaming, merepresentasikan preferensi pengguna melalui umpan balik eksplisit, seperti peringkat, atau umpan balik implisit, seperti waktu yang dihabiskan untuk menonton konten. Terlepas dari bagaimana umpan balik pengguna dikumpulkan, penggunaan data ini menimbulkan bias seleksi. Hal ini terjadi karena beberapa konten mungkin mengumpulkan sebagian besar preferensi pengguna sementara yang lain tetap berada di bagian ekor panjang item yang jarang dilihat. Jenis bias seleksi ini juga dikenal sebagai bias popularitas, di mana item yang lebih populer cenderung

dibagikan oleh lebih banyak orang dan, oleh karena itu, lebih sering direkomendasikan. Namun, popularitas tidak selalu merupakan proksi untuk kualitas konten, karena dipengaruhi oleh visibilitas item. Visibilitas dapat dimanipulasi oleh kampanye pemasaran untuk film atau lagu dan bahkan oleh bot dalam konteks kampanye politik di media sosial.

Mesin pencari juga menunjukkan bias ini karena mereka menggunakan umpan balik pengguna untuk memberi peringkat dokumen. Hal ini mengakibatkan bias peringkat. Berbagai studi telah menunjukkan bahwa dokumen di peringkat teratas lebih mungkin dipilih oleh pengguna, yang memperkuat posisi mereka seiring waktu, tetapi belum tentu karena kualitasnya yang sebenarnya. Jenis bias seleksi lainnya adalah pengambilan sampel non-acak, yang terjadi ketika data dikumpulkan dengan cara yang tidak memberikan setiap individu atau titik data dalam populasi target probabilitas yang sama untuk dimasukkan ke dalam sampel. Ketika model AI dilatih pada data yang berasal dari sampel non-acak atau non-representatif tersebut, bias dapat berkembang yang secara tidak proporsional mencerminkan karakteristik kelompok yang terlalu terwakili dalam kumpulan data. Akibatnya, kinerja model mungkin miring, yang mengarah pada hasil yang tidak akurat, tidak adil, atau tidak dapat diandalkan ketika diterapkan pada populasi yang lebih luas dan lebih beragam.

Misalnya, model penilaian kredit yang dilatih terutama pada data dari populasi yang lebih kaya dapat menilai kelayakan kredit individu dari kelompok berpenghasilan rendah secara tidak adil, sehingga melanggengkan ketimpangan ekonomi. Demikian pula dalam sistem peradilan pidana, model kepolisian prediktif yang dilatih berdasarkan data kejahatan yang biasa,yang sering kali dikumpulkan lebih banyak dari lingkungan minoritas—dapat memperkuat praktik diskriminatif, yang mengarah pada pengawasan polisi yang berlebihan terhadap komunitas ini.

Bias pengukuran

Bias pengukuran mengacu pada kesalahan sistematis yang terjadi ketika fitur atau label yang digunakan dalam model prediktif tidak secara akurat atau konsisten mencerminkan konstruk dasar yang ingin diukur. Bias ini muncul ketika data yang dikumpulkan atau dihitung sebagai proksi untuk konsep kompleks gagal menangkap cakupan penuh konstruk atau ketika metode pengukuran (atau presisinya) bervariasi di berbagai kelompok.

Bias pengukuran muncul setelah sampel dikonstruksi. Jenis bias ini terjadi ketika menyusun deskriptor untuk sampel, khususnya ketika memilih karakteristik dan variabel target yang diminati. Karakteristik deskriptif dan variabel target ini berfungsi sebagai proksi (pengukuran konkret) untuk mengaproksi entitas abstrak (ideal) yang tidak dapat dikodekan secara langsung. Biasanya, yang kita ukur adalah reduksi satu dimensi dari objek yang jauh lebih kompleks. Dalam reduksi ini, proksi gagal menangkap kompleksitas penuh dari objek yang dideskripsikannya.

Misalnya, pertimbangkan tantangan dalam memprediksi apakah seorang karyawan akan efektif dalam pekerjaannya. Konsep "efektivitas pekerjaan" memiliki banyak aspek dan tidak dapat sepenuhnya ditangkap oleh satu atribut terukur. Namun, perancang algoritma mungkin menggunakan "tahun pengalaman" sebagai proksi untuk efektivitas pekerjaan. Pendekatan ini mengabaikan faktor-faktor penting lainnya, seperti kemampuan beradaptasi,

keterampilan komunikasi, atau kreativitas, yang merupakan indikator penting kinerja pekerjaan tetapi dapat bervariasi secara signifikan di berbagai peran atau industri. Akibatnya, hanya mengandalkan "tahun pengalaman" sebagai proksi dapat menyebabkan prediksi bias yang tidak secara akurat mencerminkan potensi karyawan yang sebenarnya, terutama mereka yang berasal dari latar belakang beragam yang mungkin unggul di bidang lain.

Bias pengukuran juga dapat muncul dari disparitas dalam pengukuran proksi. Sebagai contoh, mari kita asumsikan bahwa dalam konteks pendidikan, kita menggunakan nilai sebagai deskriptor keberhasilan akademik siswa. Jika sampel mempertimbangkan dua generasi, kedua nilai tersebut tidak dapat dibandingkan secara langsung karena setiap generasi dinilai dengan instrumen evaluasi yang berbeda. Akibatnya, menarik kesimpulan dengan membandingkan nilai nilai mentah antar generasi tidaklah tepat karena skornya tidak berarti persis sama dalam tes yang berbeda.

Lebih lanjut, bias pengukuran dapat terjadi karena disparitas efektivitas pengukuran di berbagai kelompok. Misalnya, skala nyeri yang digunakan dalam layanan kesehatan merupakan pengukuran subjektif, dengan perbedaan persepsi nyeri di berbagai kelompok gender, ras, atau etnis. Karena skala persepsi nyeri bersifat subjektif, penilaian yang setara untuk individu yang berbeda dapat mewakili jenis komplikasi klinis yang berbeda, sehingga menghasilkan luaran yang keliru. Bias pengukuran juga dapat muncul dari penghilangan variabel-variabel penting. Meskipun penggunaan proksi untuk mengukur fenomena kompleks cenderung menyederhanakan realitas secara berlebihan, mengecualikan variabel relevan mengabaikan faktor-faktor eksogen yang tidak dipertimbangkan selama penyusunan deskriptor kumpulan data. Pertimbangkan contoh pembangunan sistem penilaian risiko kredit. Setelah mengumpulkan data dari nasabah perbankan, kita dapat memperoleh proksi untuk menggambarkan perilaku keuangan mereka.

Tujuannya adalah untuk menentukan apakah seorang nasabah layak kredit, dengan sistem yang memberikan peringkat risiko gagal bayar, baik tinggi maupun rendah. Bayangkan setelah model dilatih, terjadi perubahan signifikan dalam konteks ekonomi negara, seperti penyesuaian kebijakan moneter yang menyebabkan kenaikan suku bunga acuan yang ditetapkan oleh Bank Sentral.

Dalam skenario baru ini, biaya kredit meningkat, sehingga meningkatkan risiko gagal bayar. Akibatnya, kriteria pemberian kredit seharusnya menjadi lebih ketat, yang menyebabkan proporsi individu yang diklasifikasikan sebagai berisiko tinggi menjadi lebih tinggi. Namun, karena model tidak dilengkapi untuk memperhitungkan perubahan kondisi ekonomi ini, model tersebut terus mengklasifikasikan beberapa individu sebagai berisiko rendah, meskipun mereka kini lebih mungkin gagal bayar. Kelompok yang paling terpengaruh oleh bias pengukuran ini adalah mereka yang kelayakan kreditnya mendekati ambang batas keputusan, di mana pengabaian variabel-variabel penting ini memiliki dampak terbesar pada akurasi klasifikasi.

Bias Algoritmik

Bias algoritmik mengacu pada kesalahan sistematis dan berulang dan dapat muncul pada berbagai tahap siklus hidup pembelajaran mesin, termasuk pelatihan model dan proses pengambilan keputusan. Berbeda dengan kesalahan acak, bias algoritmik tertanam dalam desain dan fungsi algoritma, yang menyebabkan pola ketidakadilan yang konsisten dalam keluarannya, yang dapat membuat bias data yang ada lebih jelas.

Di antaranya, kita dapat menemukan bias yang diperkenalkan oleh algoritma pembelajaran mesin itu sendiri (bias induktif), bias yang timbul dari bagaimana model dibangun (bias agregasi), atau bahkan keputusan yang dibuat selama fase desain eksperimen yang memengaruhi pemilihan model (bias evaluasi). Sumber-sumber bias ini umumnya tercakup dalam konsep bias algoritmik, yang mengacu pada segala bentuk bias yang diperoleh atau ditekankan selama proses pelatihan model dan pemilihan model. Meskipun semua sumber ini dianggap sebagai bagian dari bias algoritmik, masing-masing memiliki karakteristik uniknya sendiri.

Meskipun penjelasan tidak menjamin orang-orang memahami alasan atau logika di balik pengambilan keputusan algoritma, perdebatan untuk interpretabilitas dan penjelasan yang lebih baik terkait erat dengan gagasan bahwa pengguna berhak untuk diberi tahu tentang proses hukum informasi mereka atau setidaknya diberikan justifikasi yang dapat dipahami atas hasil yang memengaruhi mereka.

Bias induktif

Dalam pembelajaran mesin, bias induktif mengacu pada serangkaian asumsi eksplisit atau implisit yang dibuat oleh algoritma pembelajaran untuk memungkinkan induksi, yaitu proses generalisasi dari sekumpulan observasi terbatas (data pelatihan) ke model domain yang lebih luas. Tanpa bias tersebut, induksi tidak akan mungkin dilakukan, karena observasi dapat digeneralisasi dengan berbagai cara. Jika semua generalisasi potensial diperlakukan sama, prediksi akurat untuk situasi baru tidak akan mungkin dilakukan tanpa mencerminkan pengetahuan latar belakang tentang target.

Hal ini terjadi karena algoritma pembelajaran mesin beroperasi berdasarkan prinsip mengidentifikasi pola dan keteraturan statistik dalam data. Algoritma ini menangani berbagai tugas seperti pengenalan objek, klasifikasi, pelabelan sekuens, dan bahkan pembuatan konten dengan memproses kumpulan data untuk mengenali pola-pola ini. Pola-pola ini dapat direpresentasikan dengan berbagai cara, seperti asosiasi antar variabel, pola pengelompokan, pola deskriptif atau diskriminatif, atau representasi objek. Representasi ini dapat dirancang oleh ilmuwan data atau dipelajari selama pelatihan model.

Terlepas dari bagaimana suatu algoritma merepresentasikan pola yang terdeteksi, tujuan utamanya adalah membangun fungsi yang mengekstrak deskriptor tingkat tinggi dari contoh, seperti prototipe dalam algoritma pengelompokan atau variabel target dalam model prediktif. Jika variabel target bersifat kontinu, model yang dihasilkan disebut model regresi; jika bersifat kategoris, model tersebut dikenal sebagai pengklasifikasi.

Karakteristik utama pembelajaran mesin adalah, dalam membangun fungsi yang memetakan data ke variabel target, algoritma memprioritaskan pola-pola tertentu sambil membuang pola-pola lain yang dianggap kurang signifikan. Proses pembelajaran ini dipandu

oleh fungsi yang mengevaluasi keselarasan model dengan data, baik dengan menilai kinerja model secara langsung pada suatu tugas (misalnya, menggunakan fungsi akurasi pada variabel kategoris) maupun dengan mengevaluasi kemampuan model untuk mempertahankan informasi dari kumpulan data asli dengan kerugian minimal. Dalam beberapa pendekatan, terdapat hubungan yang menguntungkan antara kinerja tugas dan retensi informasi; meminimalkan kehilangan informasi selama konstruksi model dapat secara langsung meningkatkan metrik kinerja untuk tugas-tugas tertentu.

Algoritma pembelajaran mesin secara inheren memprioritaskan pola-pola tertentu di atas pola-pola lainnya. Meskipun beberapa pola dapat membantu meminimalkan hilangnya informasi, pola lainnya mungkin diabaikan karena kurang relevan dengan hasil tugas yang diharapkan. Keputusan untuk memprioritaskan pola tertentu merupakan bias induktif seperangkat asumsi yang dipelajari model dari contoh yang memungkinkannya menyimpulkan variabel target. Namun, memprioritaskan satu tujuan di atas tujuan lainnya dapat menimbulkan tantangan, seperti memaksimalkan kinerja tugas secara keseluruhan (misalnya, akurasi klasifikasi), yang mengakibatkan kinerja yang lebih buruk bagi kelompok yang kurang terwakili. Ketidakseimbangan kelas, sebuah tantangan signifikan dalam pembelajaran mesin, memerlukan pertimbangan yang cermat karena dampaknya terhadap kewajaran dan efektivitas model.

Faktor lain yang berkontribusi terhadap bias selama pelatihan adalah efek pemangkasan. Karena model merupakan representasi data yang ringkas, penekanan pada pola tertentu di atas pola lainnya dapat memperparah disparitas bagi kelompok yang kurang terwakili dalam kumpulan data. Hal ini terjadi karena kelompok-kelompok ini lebih kecil dan, oleh karena itu, menawarkan kekayaan deskriptif yang lebih sedikit. Efek ini khususnya terlihat jelas dalam Model Bahasa Besar (LLM), sebuah topik yang akan dibahas lebih lanjut dalam bab-bab selanjutnya dalam buku ini.

Bias Agregasi

Jenis bias ini muncul ketika data digabungkan atau diagregasi sedemikian rupa sehingga mengaburkan perbedaan penting antar subkelompok dalam data. Misalnya, ketika model mengasumsikan bahwa hubungan atau pola yang diidentifikasi pada tingkat agregasi yang lebih tinggi (seperti rata-rata atau tren umum) konsisten di semua subkelompok, tetapi hal ini belum tentu demikian. Subkelompok yang berbeda dalam suatu dataset mungkin memiliki latar belakang, budaya, atau norma yang berbeda, sehingga menyebabkan variasi dalam cara menginterpretasikan variabel tertentu, yang dapat menghasilkan model yang terutama cocok untuk populasi dominan atau yang sub-optimal untuk kelompok tertentu, terutama jika dikombinasikan dengan bias representasi.

Asumsi bahwa satu model dapat menangkap keragaman suatu dataset dipertanyakan. Seringkali, dataset menggabungkan populasi yang berbeda, masing-masing menunjukkan pola yang unik. Memaksa satu model untuk mensintesis populasi heterogen tersebut untuk suatu tugas terlalu ambisius. Seperti kebanyakan pengklasifikasi konvensional (misalnya, regresi logistik atau mesin vektor pendukung), model homogen tradisional beroperasi berdasarkan asumsi model monolitik ini. Asumsi ini menyebabkan bias agregasi, di mana model, dalam

upayanya mengakomodasi keragaman dataset, dapat menggabungkan kelompok menggunakan deskriptor yang sama, sehingga mengabaikan nuansa masing-masing kelompok. Akibatnya, model menjadi sub-optimal untuk kelompok-kelompok dalam dataset alih-alih secara akurat merepresentasikan deskriptor masing-masing kelompok yang berbeda. Misalnya, bias agregasi sering terjadi dalam kumpulan data dari media sosial.

Pertimbangkan untuk membangun model klasifikasi tweet dengan variabel objektif berupa polaritas tweet, tugas umum dalam analisis media sosial untuk analisis sentimen. Asumsikan kita mengumpulkan tweet di California, AS, dari populasi berbahasa Inggris dan berbahasa Spanyol. Meskipun penggunaan emoji mungkin berbeda maknanya di antara kelompok-kelompok ini, model monolitik yang dilatih pada data ini akan menggabungkan contoh-contoh menjadi satu model, yang mengarah pada bias agregasi. Bias ini mengabaikan konteks spesifik setiap kelompok, yang berpotensi menyebabkan kesalahan klasifikasi.

Bias agregasi berkaitan dengan paradoks Simpson. Menurutnya, kesimpulan yang ditarik dari analisis populasi heterogen secara keseluruhan mungkin tidak berlaku ketika populasi dipecah menjadi strata. Ini karena mengamati tren agregat di seluruh populasi heterogen tidak memperhitungkan kekhususan setiap subkelompok. Misalnya, pertimbangkan untuk menganalisis keberhasilan akademis pada ujian masuk universitas standar. Observasi mungkin menunjukkan bias terhadap populasi Kaukasia dibandingkan populasi Pribumi, dengan orang Kaukasia memasuki program sarjana yang lebih selektif sementara pelamar Pribumi memenuhi syarat untuk mata kuliah yang kurang selektif.

Kesimpulan yang mungkin dapat disimpulkan adalah bahwa sistem seleksi universitas bias terhadap populasi Pribumi, sehingga melanggengkan ketidakadilan melalui tes standar. Namun, setelah memisahkan data, kami menemukan bahwa aplikasi Pribumi biasanya menargetkan karier teknis-profesional yang lebih pendek, sementara sebagian besar Kaukasia mendaftar ke program yang lebih panjang dengan ambang batas seleksi yang lebih tinggi. Paradoks Simpson terjadi dalam skenario ini karena populasi Pribumi umumnya mendaftar ke program dengan ambang batas penerimaan yang lebih rendah, bukan karena instrumen tersebut secara terbatas menyalurkan mereka ke dalam program tersebut.

Salah satu cara untuk mengurangi bias agregasi adalah melalui model non-monolitik. Model yang menggunakan strategi untuk bekerja dengan partisi data dan menyesuaikan model spesifik untuk setiap segmen populasi lebih baik dalam mengelola keberagaman, sehingga mengurangi efek bias agregasi. Nanti dalam buku ini, kita akan mengeksplorasi berbagai strategi pembelajaran mesin yang dapat mengurangi efek ini, dengan model ensemble yang sangat efektif dalam menangani partisi data.

Bias Evaluasi

Bias evaluasi dalam pembelajaran mesin mengacu pada kesalahan sistematis yang terjadi ketika metrik, metodologi, atau kumpulan data yang digunakan untuk mengevaluasi kinerja model tidak secara akurat atau adil mencerminkan efektivitasnya di berbagai konteks, kelompok, atau tugas, sehingga berpotensi memperburuk bias dalam model.

Jenis bias ini dapat muncul dari data atau metrik yang dipilih saat melatih atau mengevaluasi model. Misalnya, jika metrik seperti akurasi klasifikasi dipantau selama

pelatihan, kurangnya representasi kelompok tertentu yang menyebabkan ketidakseimbangan kelas akan menyebabkan akurasi lebih mengutamakan kinerja kelas mayoritas. Akibatnya, model akan lebih mengutamakan kelas tertentu daripada yang lain, sehingga menghasilkan perlakuan yang berbeda. Disparitas yang disebabkan oleh metrik ini terjadi karena pengoptimalan akurasi mengabaikan keseimbangan yang diperlukan antara dua ukuran fundamental: presisi dan perolehan kembali.

Meskipun akurasi cocok dalam konteks yang seimbang, pengklasifikasi cenderung menghasilkan disparitas kinerja antara presisi dan perolehan kembali jika terjadi ketidakseimbangan kelas. Disparitas ini berkaitan dengan ketidakseimbangan antara tingkat positif benar dan positif salah, seperti yang mungkin terjadi pada kasus COMPAS (lihat detail di Bab 1). Misalnya, untuk mengidentifikasi semua target dalam suatu kelas, kita mungkin membuat kesalahan yang meningkatkan daya ingat, tetapi jika dataset tidak seimbang, hal ini akan meningkatkan tingkat positif salah. Contoh ini menunjukkan bahwa bias evaluasi dapat berasal dari pilihan metrik evaluasi yang tidak tepat selama pelatihan model.

Defisiensi metodologis ini juga dapat muncul setelah fase pelatihan model, terutama selama tahap pemilihan model, saat mengevaluasi kinerja model pada partisi pengujian. Bias evaluasi pada tahap ini dapat terjadi karena pilihan partisi data pengujian yang tidak tepat, di mana kelompok tertentu terlalu terwakili atau kurang terwakili. Ketidakseimbangan tersebut dapat mendistorsi proses evaluasi, yang mengarah pada pemilihan model yang bias terhadap kelompok yang terlalu terwakili dalam data pengujian sementara tidak memadai dalam menilai kinerjanya pada kelompok yang kurang terwakili.

Misalnya, misalkan partisi pengujian sebagian besar terdiri dari data dari kelompok demografi tertentu, seperti individu yang lebih muda dalam aplikasi layanan kesehatan. Dalam kasus tersebut, model tersebut mungkin tampak berkinerja sangat baik selama evaluasi. Namun, kinerja ini mungkin tidak dapat digeneralisasi ke kelompok demografi lain, seperti individu yang lebih tua, yang datanya kurang terwakili dalam partisi pengujian. Akibatnya, model yang dipilih untuk diterapkan mungkin suboptimal atau bahkan merugikan ketika diterapkan pada populasi dunia nyata yang lebih luas dan lebih beragam.

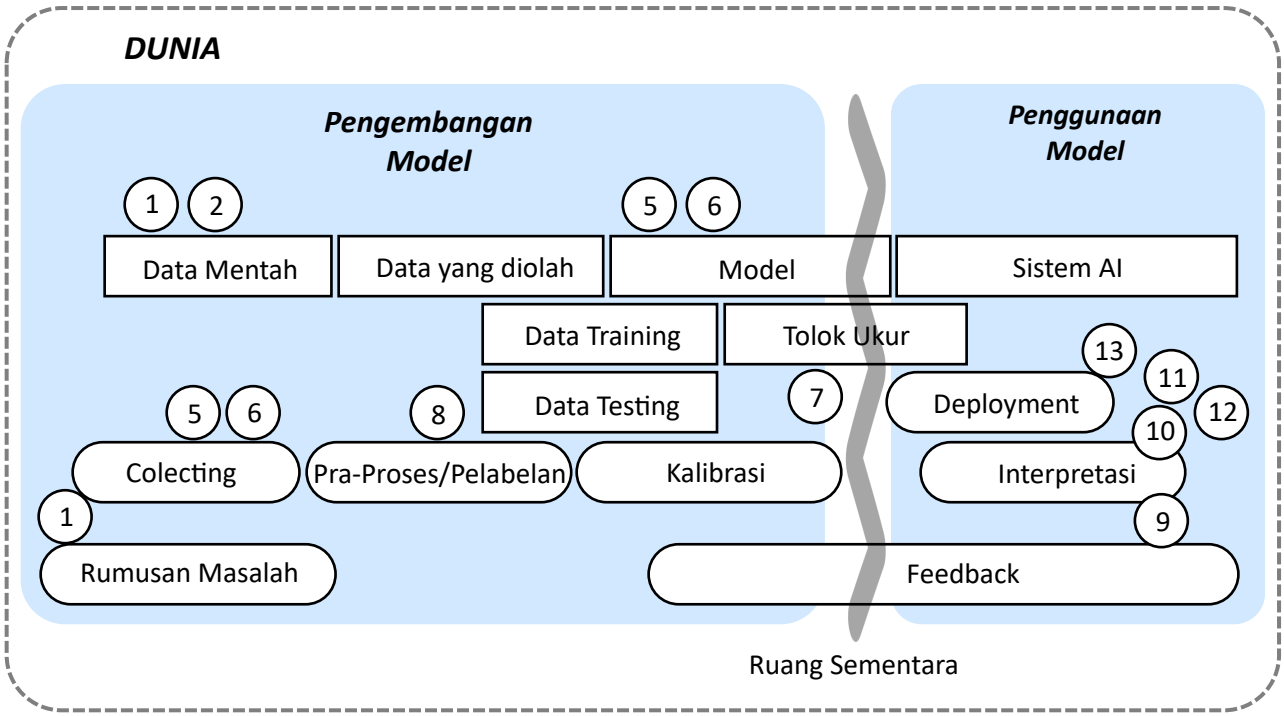
Lebih lanjut, menurut Suresh dan Guttag, suatu model dapat dioptimalkan pada data pelatihannya, tetapi kualitasnya sering diukur menggunakan tolok ukur yang mapan seperti kumpulan data UCI Machine Learning Repository, Faces in the Wild, atau ImageNet. Tolok ukur ini berfungsi sebagai standar untuk membandingkan berbagai model, yang memungkinkan evaluasi kuantitatif.

Namun, ketika tolok ukur ini tidak cukup mewakili spektrum penuh data dunia nyata, tolok ukur ini dapat mendorong pengembangan dan penerapan model yang berkinerja baik hanya pada subset data yang termasuk dalam tolok ukur. Masalah ini lebih luas daripada sumber bias lainnya karena beroperasi pada skala yang lebih luas. Tolok ukur yang keliru dapat menyebabkan adopsi model secara luas yang tampak efektif berdasarkan metrik evaluasi tetapi gagal dalam praktik ketika diterapkan pada kelompok atau skenario yang kurang terwakili.

Bias Lain dalam Pembelajaran Mesin

Sejauh ini, kami telah menjelaskan secara rinci beberapa bias paling umum dalam proses pembuatan model pembelajaran mesin. Namun, masih banyak lagi bias potensial yang dapat muncul dan memengaruhi hasil proyek yang menggunakan model pembelajaran mesin, seperti yang kami ilustrasikan pada Gambar 4.2. Misalnya, penelitian sebelumnya menunjukkan bahwa bias historis memengaruhi jenis masalah yang dipilih untuk solusi pembelajaran mesin, bahkan sebelum memilih data, variabel target, dan algoritma.

Akibatnya, banyak penggunaan AI saat ini dapat dikaitkan, misalnya, dengan pengawasan, sementara jauh lebih sedikit yang ditujukan untuk mengatasi ketidaksetaraan atau diskriminasi gender. Alur penalaran ini sejalan dengan diskusi tentang tujuan penelitian ilmiah, di mana, misalnya, masalah kesehatan yang secara eksklusif memengaruhi perempuan secara historis kurang mendapat perhatian dibandingkan penyakit lainnya. Dengan demikian, pengembang dan pengambil keputusan didorong untuk merenungkan bagaimana bias dan pengaruh sosial membentuk masalah yang ingin mereka atasi dengan AI.



Bias Data:

1. Bias Sejarah
2. Bias Representasi
3. Bias Seleksi
4. Bias Pengukuran
5. Bias Indiktif
6. Bias Agregasi
7. Bias Evaluasi
8. Bias Pelabelan
9. Bias Konfiirmasi
10. Bias Gamifikasi
11. Bias Pengetahuan Pengguna
12. Bias Otomatisasi
13. Bias yang Muncul

Gambar 4.2: Contoh bias dalam siklus hidup proyek pembelajaran mesin.

Bias lainnya melibatkan berbagai pemangku kepentingan dalam siklus hidup proyek pembelajaran mesin. Misalnya, bias dapat terjadi ketika ada kebutuhan untuk memberi label pada data atau memberikan umpan balik manusia dalam pembelajaran yang diperkuat karena

orang yang memberikan masukan dapat memasukkan bias mereka sendiri ke dalam metadata yang mereka buat.

Penting juga untuk diketahui bahwa pembuatan model pembelajaran mesin bukanlah akhir dari sebuah proyek. Model semacam itu seringkali terintegrasi ke dalam sistem dengan antarmuka pengguna tertentu, yang kemudian tertanam ke dalam proses organisasi. Sistem semacam itu dan interaksi penggunanya merupakan skenario lain di mana bias dapat muncul. Pengguna mungkin, misalnya, rentan terhadap bias otomatisasi, yang membuat mereka mengandalkan hasil otomatis tanpa pemeriksaan kritis. Alih-alih membuat daftar lengkap potensi bias lainnya di sepanjang siklus hidup pembelajaran mesin, kami bertujuan untuk mengomunikasikan bahwa ada lebih banyak bias yang perlu diperhatikan; bias-bias tersebut seringkali saling terkait dan dapat memengaruhi kualitas proyek pembelajaran mesin.

Oleh karena itu, penting untuk mempertimbangkan cara-cara mengidentifikasi, memantau, dan mengendalikan bias tersebut selama proyek AI. Kami merekomendasikan agar ini menjadi upaya kolaboratif di mana para pengembang terlibat secara aktif dan dapat menyuarakan keprihatinan mereka tentang bias. Namun, mengatasi masalah ini harus menjadi tanggung jawab bersama, yang didekati tidak hanya melalui cara teknis tetapi juga dengan strategi interdisipliner di berbagai tingkatan proses pengambilan keputusan.

Menantang Status Quo: Pendekatan Jaringan Bias (BNA)

Dalam karya mendatang dari Arriagada-Bruneau, kami mengusulkan untuk menantang metode yang berlaku dalam mengidentifikasi bias dalam proyek AI, dengan mengusulkan solusi sosioteknis. Pendekatan Jaringan Bias (BNA) adalah metode yang dirancang untuk mengatasi bias dalam pengembangan AI dengan memetakan dan memvisualisasikan interkoneksinya, alih-alih memperlakukannya sebagai kejadian yang terisolasi. Proposal ini berupaya untuk melawan apa yang kami sebut "pendekatan isolasionis", yang secara sempit berfokus pada bias individual pada tahap-tahap tertentu dari alur kerja AI (sebagaimana diilustrasikan pada Gambar 4.1 dan 4.2).

Sebaliknya, BNA bertujuan untuk mendorong refleksi etis melalui dialog terarah di antara para pengembang, yang difasilitasi oleh para ahli interdisipliner, untuk menjelaskan hubungan antara bias, sumbernya, dan dampaknya terhadap hasil proyek AI. Kami melakukan uji coba BNA dengan proyek pemrosesan bahasa alami (NLP) yang berfokus pada layanan kesehatan di Chili.

Hasilnya menunjukkan efektivitas BNA dalam mendorong transparansi, menyoroti bias yang saling terkait, dan mendorong pengembang untuk mempertimbangkan pengaruh sosial dan profesional yang lebih luas dalam pengambilan keputusan mereka. Yang terpenting, BNA mengidentifikasi keterbatasan material, faktor eksternal, dan bias profesional sebagai sumber bias yang signifikan, yang seringkali diabaikan dalam literatur bias AI tradisional. Misalnya, keterbatasan material, seperti keterbatasan sumber daya dan kualitas data yang tidak konsisten, memainkan peran penting dalam membentuk keputusan dan menimbulkan bias di seluruh proyek layanan kesehatan.

Demikian pula, bias profesional, yang berasal dari pelatihan pengembang yang berorientasi pada rekayasa, menyebabkan penekanan berlebihan pada metrik kinerja teknis

(misalnya, skor F1) dengan mengorbankan pertimbangan etika yang lebih luas, yang menggarisbawahi perlunya mengatasi bias tersebut secara lebih eksplisit dalam kerangka etika AI. Pemetaan jaringan visual BNA muncul sebagai alat penting untuk refleksi etika dan transparansi. Dengan mengilustrasikan bagaimana bias saling terkait di seluruh tahap pengembangan, alih-alih muncul sebagai contoh yang terpisah, pengembang memperoleh pemahaman yang lebih komprehensif tentang implikasi etika dari keputusan mereka. Pemetaan ini memfasilitasi diskusi kolaboratif dalam tim, memungkinkan artikulasi dan mitigasi bias yang sebelumnya tidak dikenali. Lebih lanjut, pengembang mencatat utilitas praktis visualisasi untuk meningkatkan transparansi dalam proses internal dan untuk komunikasi eksternal dengan para pemangku kepentingan seperti lembaga pemerintah. BNA bertujuan untuk melampaui kepatuhan pasif dengan daftar periksa etika dengan menanamkan refleksi etika sebagai bagian yang dinamis dan integral dari proses pengembangan AI.

Pengembang mengakui potensi metode ini untuk diterapkan pada berbagai tahap proyek AI, dari desain eksperimental hingga evaluasi retrospektif. Kemampuan adaptasi ini memposisikan BNA sebagai kerangka kerja yang berharga untuk meningkatkan kesadaran etika dan pengambilan keputusan di seluruh siklus hidup AI. Selain itu, metode sosioteknis ini sangat selaras dengan konsep imajinasi moral, yaitu kemampuan untuk mengidentifikasi dan merefleksikan secara kritis aspek-aspek etika pengambilan keputusan yang mungkin tidak langsung terlihat, serta untuk secara kreatif membayangkan perspektif dan solusi alternatif. Konsep ini, sebagaimana dibahas dalam karya Lange, merupakan kunci untuk mengatasi dilema etika yang kompleks, terutama dalam sistem sosioteknis seperti pengembangan AI karena dapat mendorong:

i. mengenali keterbatasan perspektif seseorang karena pengembang AI harus mengakui bahwa pemahaman mereka tentang suatu situasi, termasuk pilihan yang tersedia dan faktor-faktor etika yang berperan, mungkin tidak lengkap atau bias. Hal ini membutuhkan pergerakan melampaui fokus sempit yang seringkali dibentuk oleh kendala profesional atau disiplin (misalnya, memprioritaskan metrik kinerja teknis seperti akurasi atau skor F1 daripada implikasi sosial), dan

ii. mengeksplorasi perspektif alternatif secara kreatif karena pengembang didorong untuk membayangkan pendekatan atau solusi baru yang mempertimbangkan pertimbangan etika yang terabaikan, beragam sudut pandang, dan potensi dampak jangka panjang dari keputusan mereka. BNA, melalui sifat partisipatif dan reflektifnya, mengoperasionalkan imajinasi moral dengan mendorong pemikiran kritis dan kreatif tentang bias dalam AI. Dengan mendorong pengembang untuk melihat AI sebagai bagian dari jaringan interaksi manusia, masyarakat, dan teknis yang kompleks, pendekatan ini selaras dengan tujuan konsep untuk memungkinkan tata kelola antisipatif dan pengambilan keputusan yang berlandaskan etika.

Kami mendorong para pembaca yang tertarik dengan BNA untuk mengeksplorasi keterkaitan bias dan implikasi sosioteknisnya yang lebih luas, yang dapat mendukung:

- Melampaui fokus "mikroskopis" pada aspek teknis dan mempertimbangkan konteks yang lebih luas, termasuk bias sosial dan profesional.

- Antisipasi tantangan dan trade-off etika pada berbagai tahap pengembangan AI, menumbuhkan kesadaran etika yang lebih komprehensif dan mencegah komplikasi lebih lanjut.
- Mempertanyakan asumsi dan mengeksplorasi jalur alternatif, meningkatkan kemampuan untuk memperkirakan potensi bahaya atau konsekuensi yang tidak diinginkan.

BAB 5
KEADILAN, AKUNTABILITAS, DAN TRANSPARANSI AI

5.1 KEADILAN DALAM AI

Keadilan, akuntabilitas, dan transparansi merupakan prinsip-prinsip penting dalam AI sebagai sistem sosioteknis. Prinsip-prinsip ini membantu mengkarakterisasi ekosistem tempat AI dikembangkan, dengan mempertimbangkan tidak hanya tahap desain dan pengembangan tetapi juga seluruh siklus hidupnya. Bab ini membahas ketiga konsep ini, yang fundamental untuk memahami pertimbangan etis seputar AI, penggunaannya, dan apa yang nantinya akan diakui sebagai upaya regulasi.

Kita akan mulai dengan membahas konsep keadilan dalam AI. Terdapat berbagai definisi keadilan dalam literatur, dan tidak ada satu pun definisi konseptual yang dianggap dominan di bidang ini, tetapi hal ini tidak unik dalam konteks AI. Keadilan, sebagai sebuah konsep filosofis, telah lama menjadi subjek penelitian yang mendalam, sehingga sulit untuk menemukan definisi yang definitif dan diterima secara universal. Kesulitan dalam mendefinisikan keadilan berasal dari sifatnya yang kompleks, yang terkait dengan berbagai dimensi moral, sosial, dan politik. Para filsuf telah memperdebatkan esensi keadilan selama berabad-abad, tidak hanya mempertimbangkannya dalam konteks perlakuan yang setara atau keadilan, tetapi juga melalui perspektif kesetaraan, hak, dan keadilan sosial. Kompleksitas ini tercermin dalam beragam pendekatan filosofis yang mencoba merangkum arti adil masing-masing menawarkan perspektif yang seringkali bergantung pada konteks dan dipengaruhi oleh faktor budaya, sosial, dan temporal.

Misalnya, keadilan sering kali diselaraskan dengan keadilan distributif, yang berkaitan dengan alokasi sumber daya atau manfaat yang adil di antara individu, dan dikaitkan dengan karya penting John Rawls, A Theory of Justice. Baginya, konsep kuncinya didasarkan pada kerangka kerja "keadilan sebagai keadilan", yang mengusulkan prinsip perbedaan, yang secara sederhana, ia mengklaim keadilan tercapai ketika ketidaksetaraan diatur untuk menguntungkan anggota masyarakat yang paling tidak beruntung.

Melampaui keadilan distributif, keadilan meluas ke keadilan prosedural, yang mempertimbangkan keadilan proses yang mengarah pada hasil. Keadilan prosedural biasanya dinilai tanpa mempertimbangkan hasil yang dihasilkannya, melainkan menekankan pada apakah proses tersebut mengikuti aturan yang tidak memihak dan konsisten. Dimensi ini memperkenalkan kompleksitas lebih lanjut, karena keadilan harus dinilai tidak hanya dalam hal hasil tetapi juga dalam mekanisme yang digunakan untuk mencapai hasil tersebut.

Misalnya, suatu prosedur yang secara konsisten menggunakan kriteria yang bias mungkin tampak menghasilkan hasil yang "adil" ketika dievaluasi hanya berdasarkan konsistensi. Namun, konsep ini gagal memenuhi standar etika keadilan yang lebih luas, sebuah tantangan yang terlihat jelas dalam cara kita menangani bias menggunakan metrik keadilan dalam AI. Kompleksitas keadilan menjadi lebih jelas ketika kita mempertimbangkan keadilan interaksional, yang berkaitan dengan perlakuan terhadap individu dalam proses atau sistem.

Konsep ini menggarisbawahi dimensi relasional dan kontekstual dari keadilan, mengakui bahwa cara individu diperlakukan dalam interaksi interpersonal dapat sangat membentuk persepsi mereka tentang keadilan, terlepas dari faktor prosedural atau distributif. Hal ini sering dipelajari dalam lingkungan kerja dan konteks organisasi. Namun, hal ini juga dapat dipertimbangkan ketika mempelajari keadilan dalam praktik rekrutmen yang mengandalkan model AI yang dilatih dalam data historis yang bias.

Tantangan etika keadilan AI muncul justru karena sistem AI tidak netral; sistem tersebut tertanam dalam, dan sering kali memperkuat, ketidaksetaraan dan bias sosial yang ada. Penanaman ini mengharuskan pergeseran dari pendekatan keadilan yang sempit dan berbasis metrik menuju kerangka etika yang lebih holistik yang mempertimbangkan dampak AI yang lebih luas terhadap masyarakat.

Pengembang dan pemangku kepentingan lainnya dalam ekosistem AI harus menyadari risiko mereduksi keadilan menjadi serangkaian metrik kuantitatif. Meskipun metrik seperti paritas demografis atau kesempatan yang sama merupakan alat penting untuk mengidentifikasi bias dalam sistem AI, metrik tersebut pada dasarnya terbatas. Metrik-metrik ini seringkali gagal menangkap lanskap etika keadilan secara utuh, terutama ketika diterapkan tanpa mempertimbangkan konteks yang lebih luas di mana sistem AI beroperasi. Namun, hal ini tetap merupakan tugas manusia, dan merupakan tanggung jawab kita untuk secara aktif terlibat dengan kerangka kerja yang lebih luas ini untuk memikirkan keadilan dalam AI.

Dengan demikian, implikasi etis dari keadilan AI meluas ke proses dan keputusan yang membentuk sistem AI. Keadilan harus dipertimbangkan di setiap tahap siklus hidup AI, mulai dari perumusan masalah, pengumpulan data dan pengembangan model hingga penerapan, penggunaan, dan umpan balik. Pendekatan yang berorientasi pada proses terhadap keadilan ini mengakui bahwa pertimbangan etis tidak dapat dianggap sebagai renungan belakangan, tetapi harus diintegrasikan ke dalam desain dan implementasi sistem AI sejak awal. Oleh karena itu, keadilan dalam AI memerlukan kerangka kerja yang tidak hanya berfokus pada metrik atau hasil, tetapi juga meneliti proses yang mengarah pada hasil tersebut, konteks sosial yang dipengaruhinya, dan struktur kekuasaan yang berinteraksi dengannya.

Sistem AI sering kali diciptakan oleh dan untuk mereka yang berada di posisi kekuasaan, yang dapat mengarah pada penguatan hierarki sosial yang ada. Sebagaimana dikemukakan oleh Virginia Eubanks, sistem AI, ketika diterapkan dalam konteks kesejahteraan sosial, dapat melanggengkan ketimpangan struktural bahkan ketika metrik keadilan dipertimbangkan. Oleh karena itu, keadilan AI tidak dapat dipisahkan dari sistem sosial ekonomi yang berinteraksi dengan teknologi ini. Mengejar pengembangan sistem AI yang etis dan adil membutuhkan kesadaran akan dinamika kekuasaan ini dan komitmen untuk memastikan bahwa sistem AI melayani kepentingan semua orang, terutama populasi yang terpinggirkan dan rentan.

Hingga saat ini, keadilan telah dioperasionalkan dengan berbagai cara dalam AI. Operasionalisasi keadilan ini dapat dikategorikan menjadi dua kelompok tergantung pada apakah mereka menilai keadilan pada tingkat kelompok atau individu. Definisi-definisi ini didasarkan pada prinsip perlakuan yang adil, artinya definisi-definisi ini bertujuan untuk menghasilkan hasil yang serupa bagi individu atau kelompok yang serupa. Ini berarti bahwa

definisi-definisi ini berfokus pada hasil yang dihasilkan oleh model AI dan pada pendeteksian kesenjangan bagi individu atau kelompok dalam kaitannya dengan hasil-hasil ini.

Definisi Berbasis Kelompok

Sebuah model AI memenuhi konsep keadilan kelompok jika berbagai kelompok demografi dalam suatu populasi memiliki probabilitas yang sama untuk diklasifikasikan ke dalam kategori tertentu. Definisi ini menggabungkan serangkaian variabel demografi untuk analisis, yang mencakup kategori yang umum digunakan seperti gender, orientasi seksual, agama, ras, etnis, dan disabilitas. Umumnya, kami berfokus pada gender dalam definisi kami, tetapi penting untuk dicatat bahwa variabel demografi lain juga dipertimbangkan dalam konteks keadilan dalam AI.

Pertimbangkan sebuah model risiko kredit di mana variabel target Y adalah variabel biner yang menunjukkan kelayakan kredit seseorang. Model tersebut dianggap adil sehubungan dengan gender jika $P(\hat{Y} = 1\, G = m) = P(\hat{Y} = 1\, G = f)$, di mana $\hat{Y}$ adalah prediksi model dan G adalah deskriptor kelompok, dalam hal ini, gender. Gagasan utamanya adalah bahwa pelamar memiliki peluang yang sama untuk mendapatkan kredit terlepas dari gender mereka. Paritas statistik bersyarat memperluas gagasan ini dengan memasukkan serangkaian atribut yang diharapkan secara sah memengaruhi hasil.

Misalnya, dalam penugasan kredit, atribut prediktif yang sah dapat mencakup jumlah kredit yang diminta, usia pemohon, pekerjaan, dan riwayat keuangan. Faktor-faktor ini dianggap sebagai variabel kontrol dalam analisis. Dengan demikian, dalam kondisi yang serupa terkait variabel kontrol ini, pengklasifikasi tidak boleh menimbulkan disparitas dalam hasil untuk kedua kelompok. Misalkan L adalah serangkaian atribut sah yang digunakan sebagai variabel kontrol. Pengklasifikasi mempertahankan paritas statistik bersyarat untuk gender jika $P(\hat{Y} = 1|L = l, G = m) = P(\hat{Y} = 1|\, L = l, G = f)$.

Definisi ini meluas ke fase evaluasi pengklasifikasi. Gagasan utama dalam perluasan ini adalah untuk memastikan bahwa tingkat prediksi yang benar sama di berbagai kelompok. Hal ini mengarah pada gagasan paritas prediktif, yang berarti $P(Y = 1|\, \hat{Y} = 1, G = m) = P(Y = 1|\, \hat{Y} = 1, G = f)$. Jika pengklasifikasinya biner, seperti dalam penugasan kredit, paritas prediktif juga menyiratkan $P\big(Y = 0\big|\hat{Y} = 1, G = m\big) = P\big(Y = 0\big|\hat{Y} = 1, G = f\big)$. Definisi ini mencakup keseimbangan positif palsu, yang dikenal sebagai keseimbangan tingkat kesalahan positif palsu, yang didefinisikan sebagai $P\big(\hat{Y} = 1\big|Y = 0, G = m\big) = P\big(\hat{Y} = 1\big|Y = 0, G = f\big)$ dan keseimbangan tingkat kesalahan negatif palsu, yang didefinisikan sebagai $P(\hat{Y} = 0|Y = 1, G = m) = P(\hat{Y} = 0|Y = 1, G = f)$. Secara matematis, tingkat kesalahan negatif palsu menyiratkan $P(\hat{Y} = 1|Y = 1, G = m) = P(\hat{Y} = 1|Y = 1, G = f)$, suatu kondisi yang dikenal sebagai kesempatan yang sama. Ini berarti bahwa probabilitas seseorang dengan riwayat kredit yang baik untuk mendapatkan pinjaman harus sama untuk kedua kelompok.

Konsep kewajaran lainnya disebut peluang yang disamakan, juga dikenal sebagai kesetaraan akurasi prosedur bersyarat atau perlakuan yang berbeda. Peluang yang disamakan menggabungkan dua kondisi: suatu pengklasifikasi memenuhi definisi kewajaran menurut peluang yang disamakan jika kelompok yang dianalisis memiliki rasio positif benar dan positif salah yang sama. Hal ini setara dengan konjungsi definisi keseimbangan rasio kesalahan positif

salah dan keseimbangan rasio kesalahan negatif salah, yaitu, $P(\hat{Y} = 1|Y = i, G = m) = P(\hat{Y} = 1|Y = i, G = f)$, di mana $i \in \{0.1\}$.

Ide di balik definisi ini adalah bahwa baik pelamar dengan kredit baik maupun buruk harus memiliki kinerja yang sebanding dalam pengklasifikasi, terlepas dari kelompok yang dianalisis. Demikian pula, kita dapat mengukur kesetaraan akurasi penggunaan bersyarat sebagai konjungsi dari kondisi keadilan paritas prediktif, yaitu, $P\left(Y = 1\middle|\hat{Y} = 1, G = m\right) = P\left(Y = 1\middle|\hat{Y} = 1, G = f\right)$ AND $P(Y = 0|\hat{Y} = 0, G = m) = P(Y = 0|\hat{Y} = 0, G = f)$. Definisi ini menyiratkan akurasi yang setara untuk kedua kelompok di kedua kelas. Dalam contoh alokasi kredit, definisi tersebut menyiratkan bahwa probabilitas pelamar dengan kredensial baik mendapatkan kredit sama dengan probabilitas pelamar dengan kredensial buruk tidak mendapatkan kredit.

Definisi tambahan tentang keadilan dalam AI berfokus pada pengukuran kesalahan yang dibuat oleh pengklasifikasi, bukan akurasinya. Kesetaraan perlakuan didasarkan pada rasio kesalahan positif palsu dan kesalahan negatif palsu. Sebuah pengklasifikasi dianggap memiliki kesetaraan perlakuan antara dua kelompok jika rasio kesalahan mereka sama, yaitu,

$$\frac{FN}{FP}|\text{male} = \frac{FN}{FP}|\text{female}$$

Lebih lanjut, definisi kewajaran berbasis kelompok lainnya mempertimbangkan skor probabilitas yang diprediksi oleh pengklasifikasi, alih-alih label yang diprediksi $\hat{Y}$. Kewajaran pengujian, juga dikenal sebagai kalibrasi atau pencocokan frekuensi kondisional, didefinisikan sebagai paritas prediktif, yang mempertimbangkan fraksi prediksi kelas positif untuk setiap nilai s tertentu. Dalam konteks alokasi kredit, kewajaran pengujian menyiratkan bahwa $P(Y = 1|S = s, G = m) = P(Y = 1|S = s, G = f)$. Pengklasifikasi biasanya hanya memenuhi sebagian kriteria kewajaran pengujian karena kinerja pengklasifikasi cenderung menurun untuk nilai s yang mendekati ambang batas keputusan yaitu, $s \rightarrow 0{,}5$. Hal ini terjadi karena contoh-contoh yang berada di dekat 'zona kebingungan' antarkelas cenderung mengumpulkan lebih banyak contoh yang salah klasifikasi.

Definisi Berbasis Individu

Dalam konteks mendefinisikan atribut yang sah yang digunakan dalam definisi berbasis kelompok seperti paritas statistik bersyarat, definisi berbasis individu bergantung pada kesadaran atau ketidaksadaran akan atribut sensitif. Definisi-definisi ini didasarkan pada identifikasi serangkaian atribut sensitif yang tidak boleh dipertimbangkan oleh pengklasifikasi ketika memprediksi variabel target. Konsekuensi utama dari pendekatan ini adalah bahwa dua individu dengan atribut non-sensitif yang identik seharusnya menerima hasil yang sama dari pengklasifikasi.

Konsep ini mengarah pada apa yang kami sebut "keadilan melalui ketidaksadaran." Sebuah pengklasifikasi dianggap tidak sadar jika tidak menggunakan atribut sensitif untuk menghasilkan hasilnya. Misalnya, dalam pemberian kredit, sebuah pengklasifikasi tidak menyadari atribut sensitif gender jika atribut ini tidak dipertimbangkan selama proses pelatihan. Keadilan melalui ketidaksadaran menyiratkan bahwa dua subjek, i dan j, yang memiliki atribut non-sensitif yang sama seharusnya menerima hasil yang sama; yaitu, jika

$X_i = X_j$, maka $\hat{Y}_i = \hat{Y}_j$. Hasil ini juga menunjukkan bahwa kumpulan data tidak menggunakan proksi gender untuk menghasilkan hasilnya.

Definisi lain tentang kewajaran berdasarkan individu disebut kewajaran melalui kesadaran, yang menangkap prinsip bahwa individu yang serupa harus menerima hasil yang identik. Dalam konteks ini, kesamaan individu diukur dengan metrik jarak. Jika pengklasifikasi adil, jarak antara distribusi keluaran individu tidak boleh lebih besar dari jarak antara representasi mereka.

Secara formal, untuk sekumpulan individu U, metrik jarak individu $d: U \times U \to \mathbb{R}$, pemetaan dari U ke distribusi probabilitas atas hasil model $M: U \to P$, dan metrik jarak distribusional $D: P \times P \to \mathbb{R}$, kita katakan bahwa pengklasifikasi adil melalui kesadaran untuk dua individu i, j jika dan hanya jika $D(M(i), M(j)) \leq d(i, j)$. Misalnya, pertimbangkan dua subjek i, j yang representasinya X_i dan X_j memiliki jarak Euclidean 0,3. Model risiko kredit menghasilkan probabilitas berikut: $P(\hat{Y}_i = 1|X_i) = 0{,}6$ dan $P(\hat{Y}_j = 1|X_j) = 0{,}7$. Secara distribusional, mengingat sifat biner pengklasifikasi, subjek menerima probabilitas $M(i) = [0{,}6, 0{,}4]$ dan $M(j) = [0{,}7, 0{,}3]$. Metrik probabilitas yang umum digunakan adalah jarak statistik, juga dikenal sebagai variasi total, dilambangkan sebagai $D_{tv}(P, Q) = \frac{1}{2}\sum |P(a) - Q(a)|$, yang sesuai dengan setengah dari norma L1 (selisih absolut). Dalam kasus kami, $D_{tv} = 0{,}1$. Karena $D_{tv}(M(i), M(j)) = 0{,}1 \leq d(X_i, X_j) = 0{,}3$, kami menyatakan bahwa pengklasifikasi tersebut adil melalui kesadaran untuk individu i dan j.

5.2 PENALARAN KAUSAL

Definisi berbasis kelompok dan berbasis individu beroperasi pada tingkat yang berbeda. Definisi berbasis kelompok menyoroti pentingnya karakteristik demografis dan, berdasarkan karakteristik tersebut, mendefinisikan kondisi keadilan di antara kelompok, sementara definisi berbasis individu berfokus pada penggunaan atau non-penggunaan atribut yang dilindungi, seperti keanggotaan atau identifikasi individu dalam suatu kelompok. Baik definisi berbasis kelompok maupun berbasis individu bergantung pada penggunaan atribut sensitif selama konstruksi model atau pada kemungkinan mengidentifikasi perbedaan dalam keluaran model berdasarkan atribut-atribut ini.

Pendekatan alternatif melibatkan penilaian apakah atribut-atribut ini benar-benar berpengaruh pada keluaran model. Pendekatan ini mengarah pada analisis kausalitas, yang mencakup eksplorasi apakah memang terdapat bukti yang mendukung bahwa suatu atribut tertentu menyebabkan keluaran tertentu. Definisi keadilan berdasarkan penalaran kausal didasarkan pada prinsip bahwa suatu keputusan dianggap adil bagi seorang individu jika keputusan tersebut tetap sama dalam kondisi saat ini atau dalam kondisi alternatif yang mengakibatkan perubahan pada atribut yang dilindungi. Misalnya, sebuah pengklasifikasi dianggap adil dalam penalaran kausal terkait atribut gender yang dilindungi bagi seorang individu jika hasil pengklasifikasi tersebut akan sama terlepas dari gendernya.

Dalam penalaran kausal, ketergantungan antar variabel diformalkan menggunakan graf berarah. Konsep graf ketergantungan berarah melibatkan pembentukan hubungan

ketergantungan kausal antar variabel. Kausalitas tidak sama dengan korelasi. Dalam hubungan kausal, hubungan diarahkan dari preseden (sebab) ke konsekuen (akibat).

Cara umum untuk menganalisis model dari perspektif penalaran kausal adalah dengan mendefinisikan graf kausal antar variabel. Graf kausal adalah graf asiklik berarah di mana simpul merepresentasikan variabel dan sisi merepresentasikan hubungan antar variabel tersebut. Graf kausal digunakan untuk menganalisis pengklasifikasi.

Graf kausal mempertimbangkan berbagai jenis simpul. Sebuah simpul merupakan proksi jika nilai variabel yang diwakilinya dapat digunakan untuk menentukan nilai variabel lain dalam graf tersebut. Jenis hubungan antara variabel proksi dan variabel turunan ini direpresentasikan dalam grafik kausal dengan tepi berarah dari simpul turunan ke simpul proksi. Misalnya, dalam kasus alokasi kredit, mari kita asumsikan kita memiliki variabel Boolean yang disebut 'Sindrom Turner', yang menunjukkan apakah orang yang mengajukan kredit memiliki sindrom ini. Sindrom Turner adalah kelainan genetik yang memengaruhi perkembangan seksual dan reproduksi anak perempuan. Karena hanya dapat memengaruhi perempuan, variabel 'Sindrom Turner' merupakan proksi untuk gender. Grafik kausal akan merepresentasikan hubungan ini dengan tepi berarah dari 'Gender' ke 'Sindrom Turner'.

Hubungan kausal antara variabel dan proksinya harus dipertimbangkan secara cermat jika variabel independen dilindungi. Misalnya, jika gender merupakan variabel yang dilindungi, penggunaan 'Sindrom Turner' sebagai proksi, meskipun variabel gender itu sendiri tidak digunakan dalam pengklasifikasi, akan menghasilkan hasil yang bias gender, sehingga membuat model menjadi tidak adil terhadap variabel yang dilindungi. Jika hubungan antara variabel terlindungi dan variabel dependen tidak mengubah hasil pengklasifikasi, variabel dependen tersebut disebut atribut penyelesaian.

Mari kita pertimbangkan variabel 'Jumlah Kredit', yang bergantung pada atribut terlindungi 'Jenis Kelamin'. Ketergantungan ini muncul karena, rata-rata, perempuan mengajukan pinjaman yang lebih kecil daripada laki-laki. Namun, jika kita berasumsi bahwa 'Jumlah Kredit' tidak memengaruhi keputusan pengklasifikasi, artinya kredit diberikan untuk jumlah rendah dan tinggi, maka 'Jumlah Kredit' akan dianggap sebagai atribut penyelesaian untuk 'Jenis Kelamin'.

Grafik kausal dapat mencakup berbagai jenis hubungan antar variabel. Dalam konteks analisis kewajaran, hubungan ini hanya menarik jika melibatkan atribut terlindungi atau atribut proksi. Dari analisis grafik kausal, kita memiliki dua pendekatan terhadap kewajaran dalam pembelajaran mesin (ML) berdasarkan penalaran kausal. Kita mengatakan bahwa pengklasifikasi bersifat kontrafaktual adil jika hasilnya tidak bergantung pada atribut terlindungi. Misalnya, sebuah pengklasifikasi tidak adil secara kontrafaktual jika menggunakan variabel 'Sindrom Turner'.

Meskipun pengklasifikasi tidak menggunakan atribut yang dilindungi 'Gender' (sehingga mencapai keadilan melalui ketidaksadaran), pengklasifikasi tersebut gagal menjadi adil secara kontrafaktual karena menggunakan 'Sindrom Turner', yang merupakan proksi untuk 'Gender'. Kami mendefinisikan sebuah pengklasifikasi sebagai bebas dari diskriminasi proksi

jika tidak ada jalur dari atribut yang dilindungi ke keluaran pengklasifikasi yang dimediasi oleh proksi.

Pengklasifikasi mungkin bebas dari diskriminasi proksi tetapi tidak adil secara kontrafaktual. Situasi ini terjadi jika pengklasifikasi tidak menggunakan variabel 'Sindrom Turner' tetapi menggunakan 'Jenis Kelamin'. Dalam hal ini, tidak ada proksi untuk atribut yang dilindungi, tetapi pengklasifikasi tidak adil karena menggunakan variabel yang dilindungi 'Jenis Kelamin'. Penting untuk dicatat bahwa adil secara kontrafaktual dan diskriminasi proksi adalah dua definisi keadilan yang saling melengkapi dalam Pembelajaran Mesin (ML). Namun, contoh tersebut menunjukkan bahwa diskriminasi proksi adalah definisi keadilan yang lebih lemah daripada adil secara kontrafaktual. Hal ini karena suatu model dapat memenuhi definisi diskriminasi proksi tetapi gagal memenuhi definisi adil secara kontrafaktual.

Secara keseluruhan, terdapat banyak cara untuk mengukur dan menilai keadilan dalam AI, tetapi tantangannya terletak pada penentuan mekanisme mana yang paling sesuai dengan konseptualisasi keadilan spesifik yang dibutuhkan untuk suatu proyek. Penelitian sebelumnya menunjukkan bahwa secara matematis mustahil untuk memenuhi berbagai definisi metrik keadilan yang diukur secara komputasi secara bersamaan. Oleh karena itu, kerangka kerja keadilan AI telah mengusulkan pedoman tentang metrik mana yang harus diprioritaskan berdasarkan jenis prediksi yang dibuat oleh setiap AI tertentu. Meskipun demikian, bidang penelitian dan praktik ini terus berkembang secara aktif dan kami berharap akan melihat perkembangan konseptual dan teknis baru dalam beberapa tahun mendatang, terutama dalam melampaui AI prediktif dan mengatasi tantangan AI generatif.

5.3 AKUNTABILITAS DALAM AI

Meskipun transparansi krusial untuk memahami bagaimana sistem AI mengambil keputusan, konsep akuntabilitas membawa kita untuk berpikir tentang bagaimana sistem ini bertanggung jawab atas hasil dan tindakannya. Pengembangan sistem cerdas menawarkan peluang untuk meningkatkan efisiensi berbagai aktivitas, tetapi juga menghadirkan skenario baru yang melibatkan konsekuensi tak terduga dan tantangan etika mendasar.

Akuntabilitas krusial karena sistem cerdas memungkinkan kita mendelegasikan tugas kepada algoritma. Meskipun terdapat banyak definisi akuntabilitas, terutama dalam dokumen terkait tata kelola AI seperti GDPR atau ALTAI, semuanya sepakat pada satu aspek: akuntabilitas adalah kewajiban untuk melaporkan dan membenarkan tindakan kepada otoritas. Mengikuti Bovens, akuntabilitas adalah konsep umum yang mencakup dan tumpang tindih dengan berbagai gagasan seperti transparansi dan tanggung jawab. Pada intinya, akuntabilitas melibatkan hubungan antara aktor dan forum, di mana aktor harus menjelaskan dan membenarkan tindakannya, forum dapat mengajukan pertanyaan dan menilai, dan akibatnya, aktor tersebut dapat menghadapi konsekuensi. Aktornya bisa berupa individu, tetapi bisa juga organisasi (publik atau swasta).

Di sisi lain, forum bisa berupa individu yang diberi wewenang, biasanya berkonotasi publik, seperti jurnalis, hakim, atau jaksa. Penting untuk dicatat bahwa konotasi publik tidak selalu berkaitan dengan pemerintah. Kewenangan publik mungkin signifikan karena individu

tersebut menjalankan peran publik, seperti memberi informasi. Dalam skenario ini, apakah individu tersebut berasal dari pemerintah atau entitas swasta tidaklah relevan. Individu tersebut diberi wewenang melalui konotasi publik karena perannya. Forum bisa berupa entitas negara, seperti pengadilan, atau badan pemerintah, seperti badan regulasi atau pengawasan.

Kewajiban aktor untuk melapor kepada forum dapat muncul dalam konteks hukum dan informasi, yang tidak saling eksklusif. Instansi informasi bertujuan untuk mengumpulkan informasi sensitif berdasarkan kesaksian aktor, sementara tanggung jawab hukum yang timbul dari informasi ini nantinya dapat berada di bawah instansi hukum yang dapat melibatkan potensi sanksi.

Akuntabilitas memiliki berbagai bentuk, tergantung pada aktor dan forum yang terlibat. Terdapat akuntabilitas politik, yang menjadi tanggung jawab pejabat terpilih. Terdapat pula akuntabilitas hukum, yang berlaku bagi warga negara dalam kerangka hukum dan organisasi hukum privat. Baik dalam akuntabilitas politik maupun hukum, forum dapat menjatuhkan sanksi formal sebagaimana didefinisikan dalam kerangka hukum masing-masing negara. Selain itu, terdapat akuntabilitas administratif, di mana forum terdiri dari auditor dan regulator yang bertindak secara independen. Jenis akuntabilitas ini berlaku untuk pengawasan keuangan, pengendalian proses eksternal, dan lembaga akreditasi mutu. Aktor dalam jenis hubungan ini umumnya adalah organisasi publik atau swasta.

Terdapat pula akuntabilitas profesional, di mana aktornya adalah para profesional, dan forumnya adalah asosiasi dan serikat profesi yang mengawasi praktik etis profesi mereka. Terakhir, terdapat akuntabilitas sosial yang dipicu oleh meningkatnya ketidakpercayaan terhadap lembaga negara dan peran regulasinya. Dengan tidak adanya forum regulasi independen, organisasi non-pemerintah seringkali meluncurkan inisiatif yang dikelola sendiri untuk memenuhi peran ini. Karena lembaga-lembaga ini tidak memiliki kewenangan sanksi dalam kerangka hukum, mereka tidak menjatuhkan sanksi hukum atau administratif, melainkan sanksi moral. Namun, forum-forum semacam ini nantinya dapat beralih ke lembaga peradilan jika terdapat kecurigaan adanya tanggung jawab hukum.

Dalam konteks AI, terdapat beberapa jenis hubungan akuntabilitas. Yang paling jelas adalah akuntabilitas hukum, karena perancang, pengembang, dan pejabat perusahaan yang menerapkan dan mempopulerkan sistem cerdas harus mematuhi kerangka hukum dan peraturan yang telah ditetapkan. Jika kerangka peraturan terkait aspek ini longgar, akuntabilitas sosial berperan, yang berarti masyarakat sipil harus melakukan pengawasan. Contoh-contoh kuasi-hukum ini dapat mengungkapkan informasi dan menetapkan sanksi sosial. Akuntabilitas politik dan profesional juga dapat muncul ketika AI, seperti AI generatif, digunakan untuk propaganda.

Lebih lanjut, hambatan utama bagi akuntabilitas AI adalah kurangnya transparansi yang melingkupi sistem AI. Model AI yang tidak transparan mengaburkan faktor-faktor yang berkontribusi terhadap hasil yang tidak adil atau bentuk kerugian lainnya. Misalnya, hasil ini mungkin bergantung pada data yang digunakan untuk melatih model AI, tetapi juga dapat dipengaruhi oleh algoritma pembelajaran, yang dapat memperburuk atau mendistorsi pola tertentu. Karena penerapan sistem AI dalam pengaturan organisasi pada dasarnya

mendelegasikan tugas kepada AI, dalam hal akuntabilitas, tanggung jawab menjadi terdilusi di antara berbagai faktor yang dapat menyebabkan hasil yang tidak diinginkan. Menelusuri penyebab hasil tersebut dalam sistem yang tidak transparan tetap menjadi tugas yang menantang.

Apa yang mencirikan akuntabilitas dalam AI?

Menurut Novelli, karakteristik utama akuntabilitas dalam AI meliputi konteks, cakupan, dan agen yang terlibat. Konteks mengacu pada bidang di mana AI digunakan dan tingkat otonomi sistem AI. Cakupan melibatkan tahapan spesifik dari suatu proses penerapan AI, seperti desain, pengembangan, atau penerapan. Desain melibatkan tugas-tugas seperti perencanaan, desain arsitektur, pemilihan teknologi atau model dasar, desain antarmuka, penggunaan data, dan strategi pengembangan yang digunakan. Pengembangan berkaitan dengan komposisi tim, yang dapat mencakup pemrogram, insinyur, penguji, atau koordinator tim. Penerapan melibatkan pemantauan apakah sistem memberikan hasil yang diharapkan dan pemeliharaan sistem AI tersebut. Karakteristik ketiga, agen, dapat diidentifikasi secara individual, korporat, kolektif, atau hierarkis.

Konfigurasi ini krusial untuk memahami jenis akuntabilitas AI yang dimaksud. Akuntabilitas AI dapat bersifat reaktif atau proaktif. Akuntabilitas proaktif dipandang sebagai suatu keutamaan proses, yang terintegrasi dalam tujuan perencanaan untuk mengantisipasi kejadian dan mencegah kegagalan. Sebaliknya, akuntabilitas reaktif terjadi setelah peristiwa berbahaya terjadi, yang bertujuan untuk memitigasi kegagalan yang telah terjadi.

Mekanisme akuntabilitas AI dapat melibatkan rekomendasi, persetujuan, larangan, atau berbagai jenis sanksi. Berbagai negara dan organisasi sedang berupaya mengembangkan standar akuntabilitas, dengan banyak di antaranya bertujuan untuk cakupan global. Upaya ini melibatkan negara-negara yang mematuhi perjanjian internasional atau perjanjian multilateral yang mencakup beragam bidang, termasuk aplikasi teknologi dan standar komersial. Penerapan standar-standar ini memerlukan proses yang telah ditentukan seperti audit internal, penilaian mandiri, penilaian sejawat, atau evaluasi eksternal, yang semuanya terkait dengan pemantauan efektivitas sistem dan kepatuhannya terhadap standar.

Dalam konteks ini, peran negara-bangsa sangatlah krusial. Negara-bangsa, melalui berbagai organisasi pemerintahannya, bertanggung jawab untuk mengatur ranah ini dengan mengembangkan kerangka kerja yang mendorong pengembangan AI sekaligus melindungi masyarakat sipil dari risiko terkait dan potensi skenario buruk. Upaya-upaya ini disebut sebagai tata kelola AI. Meskipun tata kelola AI sangat penting bagi pengembangan AI yang bertanggung jawab, kemajuan di bidang ini masih dalam tahap awal. Tidak ada definisi yang jelas tentang apa yang dimaksud dengan tata kelola AI atau cakupannya. Inti dari perbedaan pandangan tentang tata kelola AI bergantung pada persepsi kita tentang peran Negara. Sementara beberapa perspektif menganjurkan untuk meminimalkan peran Negara dalam pengembangan teknologi, membiarkan bidang tersebut terbuka bagi perusahaan TI transnasional besar, yang lain menyerukan pengembangan kebijakan publik dalam kerangka regulasi yang didefinisikan dengan jelas. Isu-isu seperti privasi data, implikasi etis penggunaan AI dalam pengambilan

keputusan, atau penggantian tenaga kerja merupakan area kunci untuk dibahas lebih lanjut dalam ranah ini.

5.4 TRANSPARANSI DALAM AI

Selain keadilan, transparansi merupakan aspek krusial lain dari etika AI. AI, khususnya pembelajaran mendalam, menghadapi pengawasan karena kurangnya transparansi terkait cara pengambilan keputusan atau menghasilkan konten. Hal ini sering menyebabkan AI digambarkan sebagai "kotak hitam". Isu-isu transparansi memperburuk kekhawatiran tentang hasil yang tidak adil, karena ketidakjelasan model AI membuat pendeteksian dan penanganan masalah ini menjadi lebih sulit. Selain keadilan, isu-isu transparansi juga mencakup pertimbangan etis lainnya. Misalnya, para peneliti telah menyuarakan kekhawatiran mengenai penggunaan data pribadi secara luas tanpa adanya informasi yang jelas dan mudah diakses mengenai tujuan penggunaannya dan dampak lingkungan yang besar yang terkait dengan teknologi AI.

Secara konseptual, transparansi terkait dengan gagasan pemahaman. Menurut Michael Reddy, pemahaman bergantung pada serangkaian operasi sistematis yang bergerak dari ranah objek fisik ke ranah operasi mental. Pemahaman sangat terkait dengan konsep melihat dan mengetahui, yang keduanya berkaitan dengan pemahaman. Lebih lanjut, istilah-istilah seperti menerangi, mengklarifikasi, dan membuat transparan terkait dengan tindakan pemahaman. Apa yang tidak transparan dan karenanya tidak jelas atau buram dianggap tidak dapat dipahami. Dengan demikian, tujuan utama transparansi AI adalah membuat AI dapat dipahami.

Dalam ranah AI, transparansi dioperasionalkan dalam beberapa cara. Satu perspektif menghubungkan transparansi dengan konsep keterbukaan. Sistem yang transparan dapat menjadi sistem terbuka. Sistem cerdas dianggap terbuka jika melibatkan data terbuka, sumber terbuka, dan akses terbuka. Keterbukaan suatu sistem ditentukan oleh model bisnis perusahaan yang mengembangkannya atau prinsip-prinsip yang menginspirasi komunitas pengembang. Misalnya, komunitas yang mengadvokasi data terbuka atau sumber terbuka mendorong praktik-praktik yang mendorong perancangan dan pengembangan sistem cerdas terbuka. Keterbukaan mendorong kondisi yang kondusif bagi hasil yang dapat direproduksi. Dengan demikian, reproduktifitas dan keterbukaan saling terkait.

Sebaliknya, sistem proprietary, yang menggunakan lisensi proprietary, cenderung menjadi sistem tertutup. Lebih sulit untuk memahami aturan, pola, dan model yang digunakan oleh sistem AI tersebut untuk menghasilkan hasil. Meskipun tidak semua sistem proprietary sepenuhnya tertutup, mereka biasanya menunjukkan keterbukaan yang terbatas. Misalnya, sistem AI proprietary mungkin mengungkapkan data tempat mereka dilatih, tetapi mengakses kode mereka biasanya lebih menantang. Sistem terbuka dan tertutup dapat hidup berdampingan pada platform yang sama. Misalnya, HuggingFace, sebuah platform yang terkenal untuk reproduktifitas, mendukung kedua jenis inisiatif tersebut. Dalam konteks ini, sistem tertutup dapat memiliki fitur yang meningkatkan reproduktifitas.

Namun, hubungan antara sistem sumber terbuka dan reproduktifitas terbukti lebih kuat. Pada HuggingFace, sistem tertutup menyediakan file yang dapat dieksekusi sedemikian rupa sehingga bertindak sebagai kotak hitam, memungkinkan kita untuk mengoperasikan model-model ini pada platform dan memfasilitasi reproduktifitas hasil. Perspektif krusial lain yang terkait dengan transparansi adalah keterjelasan. Keterjelasan suatu sistem merepresentasikan kemampuan untuk mengekstrak penjelasan dari model AI. Keterjelasan mengarah pada apa yang disebut Kecerdasan Buatan yang Dapat Dijelaskan (XAI), yang terdiri dari serangkaian metode yang memungkinkan model kotak hitam menghasilkan penjelasan yang meningkatkan transparansi. Kami akan membahas kembali mekanisme untuk menghasilkan penjelasan dari model AI nanti di buku ini; namun, kami akan membahas di sini peran keterjelasan dalam mencapai transparansi.

Tidak seperti keterbukaan, perspektif ini bertujuan untuk mengatasi ketidakjelasan model AI yang menghambat pemahaman proses yang mendorong keluaran tertentu. Masalah ini khususnya bermasalah untuk sistem yang menggunakan pembelajaran mendalam. Model-model ini sering berfungsi sebagai kotak hitam, yang memungkinkan pengembang untuk memverifikasi apakah mereka menghasilkan hasil yang diharapkan untuk masukan tertentu (menggunakan data uji) tetapi tidak untuk menjelaskan mekanisme internal di balik hasil ini. Akibatnya, bahkan pengembang tidak dapat sepenuhnya memahami dan mengomunikasikan alasan di balik keputusan AI ini kepada pengguna. Kurangnya transparansi ini mencegah pengguna menilai apakah keputusan AI dicapai melalui metode yang rasional atau tepat, dan hal ini menghalangi kemampuan mereka untuk mengajukan banding yang beralasan terhadap keputusan tersebut atau untuk merancang strategi guna mengamankan hasil yang lebih menguntungkan.

Mekanisme XAI bertujuan untuk menawarkan penjelasan global yang berlaku untuk semua hasil model AI atau penjelasan lokal yang disesuaikan dengan hasil tertentu. Namun, XAI terutama dirancang oleh dan untuk pengembang, seringkali mengabaikan perspektif pengguna akhir dan mereka yang terdampak oleh AI. Masih diperlukan metode standar dan efektif untuk memberi tahu pengguna non-teknis tentang AI dan fungsinya. Bidang yang lebih baru yang disebut AI yang dapat dijelaskan yang berpusat pada manusia (HCXAI) telah mengalihkan penekanan dari aspek teknis dalam menghasilkan informasi tentang model AI ke tantangan dalam mengomunikasikan informasi ini secara efektif kepada individu yang perlu memahami dan menggunakannya untuk pengambilan keputusan. HCXAI konsisten dengan konsep "transparansi yang bermakna", yang berupaya memastikan bahwa individu memahami keputusan AI dengan cara yang relevan bagi mereka. Pemahaman ini akan memungkinkan mereka untuk melakukan intervensi, menyetujui atau menolak keputusan, dan meminta pertanggungjawaban AI. Literatur mengidentifikasi beberapa kelompok yang membutuhkan penjelasan tentang model AI, termasuk pengembang AI, pakar domain (seperti hakim yang menerima skor residivisme yang diprediksi), individu yang terdampak (seperti terdakwa), pengguna umum, auditor AI, dan pembuat kebijakan.

Sejauh ini, penelitian HCXAI berfokus pada evaluasi apakah orang memahami penjelasan, menganggapnya bermanfaat, mengembangkan tingkat kepercayaan yang

terkalibrasi dengan tepat terhadap AI, dan apakah penjelasan meningkatkan kolaborasi AI-manusia. Bukti menunjukkan bahwa meskipun penjelasan AI sering kali meningkatkan pemahaman subjektif pengguna, dampaknya terhadap pemahaman dan kepercayaan objektif beragam. Beberapa penjelasan gagal mendorong penerimaan AI dan terkadang mengakibatkan kepercayaan yang tidak beralasan dan ketergantungan yang berlebihan.

Dengan demikian, dan agak tidak terduga, para peneliti telah menemukan risiko lain yang terkait dengan keterjelasan AI. Ehsan dan Riedl membahas "jebakan keterjelasan", di mana penjelasan AI dapat menyebabkan pengguna terlalu bergantung pada keputusan AI dengan mengorbankan penilaian mereka sendiri. HCXAI mengalami kemajuan, dan kami mengantisipasi peningkatan penelitian di bidang ini yang akan melengkapi kemajuan dalam XAI. Namun, kesenjangan utama yang perlu diatasi adalah bahwa sebagian besar penelitian terkonsentrasi pada konteks di Global Utara, yang menyebabkan kurangnya pemahaman tentang bagaimana teknik XAI dikembangkan, diimplementasikan, atau dinilai di komunitas di Global Selatan.

Terakhir, ada berbagai pandangan lain tentang transparansi. Beberapa pendekatan bertujuan untuk meningkatkan transparansi keputusan yang dibuat selama proses pelatihan dan evaluasi AI. Misalnya, mekanisme tertentu berfokus pada peningkatan ketertelusuran terkait pilihan data dan model, seperti mendokumentasikan keputusan dalam lembar data untuk kumpulan data dan kartu model. Metode lain melibatkan pelaksanaan audit algoritmik untuk mengevaluasi kewajaran model AI dan untuk meneliti proses yang mengarah pada luaran AI. Upaya signifikan masih diperlukan untuk menerapkan transparansi dalam proyek AI, dengan menangani aspek-aspek seperti proses, pengguna, model, dan data. Pendekatan yang menjanjikan untuk membedakan antara aspek-aspek transparansi ini ditawarkan oleh konseptualisasi transparansi berdasarkan desain, yang mengidentifikasi tiga tingkat transparansi: transparansi berdasarkan kebajikan, transparansi relasional, dan transparansi sistemik.

Transparansi berdasarkan prinsip mengacu pada tindakan pengungkapan informasi tentang operasi internal sistem AI, di mana kita dapat menempatkan sebagian besar penelitian XAI dan upaya lain untuk mempublikasikan dokumentasi dan evaluasi model. Transparansi relasional berfokus pada bagaimana orang mempersepsi dan memahami informasi ini, menyoroti bahwa terdapat berbagai jenis pengguna transparansi dan upaya perlu dilakukan untuk menilai bagaimana kebutuhan informasi mereka terpenuhi. Dengan demikian, HCXAI berada di level kedua ini. Terakhir, transparansi sistemik melibatkan konteks kelembagaan di mana terdapat hubungan antara sistem AI dan para pemangku kepentingan transparansinya. Kami mengharapkan kemajuan substansial di bidang-bidang ini di masa mendatang seiring dengan semakin banyaknya peraturan dan rekomendasi yang ada yang berfokus pada arah ini.

BAB 6
INISIATIF REGULASI DALAM AI

6.1 PENDAHULUAN

Beberapa inisiatif multilateral telah membahas implikasi etis AI. Di Eropa, inisiatif yang paling banyak dipublikasikan berasal dari badan pengambil keputusan politik, seperti Parlemen Eropa. Sebaliknya, inisiatif lain berasal dari sektor korporasi swasta, termasuk yang didefinisikan oleh perusahaan seperti Amazon, Microsoft, META, Google, dan OpenAI. Meskipun terdapat kesamaan dalam definisi dan prinsip yang ingin dijunjung tinggi oleh inisiatif-inisiatif ini, terdapat pula nuansa penting.

Inisiatif Parlemen Eropa umumnya sejalan dengan konsep AI yang Tepercaya, sementara inisiatif korporasi swasta sering merujuk pada gagasan AI yang Bertanggung Jawab. Meskipun konsep-konsep ini tidak saling eksklusif, konsep-konsep ini menekankan aspek dan nuansa berbeda yang mengungkapkan prioritas spesifik masing-masing korporasi. Kami akan meninjau aspek-aspek kunci dari beberapa inisiatif ini untuk menjelaskan perbedaan dan persamaan di antara pendekatan-pendekatan ini.

Peraturan Perlindungan Data Umum (GDPR, 2016)

GDPR (Peraturan Perlindungan Data Umum) adalah inisiatif regulasi oleh Parlemen Eropa dan Dewan Uni Eropa yang bertujuan untuk melindungi orang perseorangan terkait pemrosesan data pribadi dan pergerakan bebas data tersebut. Ditetapkan pada tahun 2016, GDPR menetapkan bahwa perlindungan data pribadi merupakan hak asasi. Aspek ini krusial untuk mengembangkan sistem cerdas, karena model AI seringkali bergantung pada data pribadi. GDPR menyediakan kerangka regulasi untuk penggunaan data pribadi di berbagai bidang, termasuk yang melibatkan pembuatan model AI.

GDPR mengakui perlindungan data pribadi sebagai hak, tetapi menyatakan bahwa hal tersebut bukanlah hak yang mutlak. Menurutnya, perlindungan data harus dirancang untuk melayani kemanusiaan. Oleh karena itu, perlindungan data harus dipertimbangkan dalam kaitannya dengan fungsi sosialnya dan diseimbangkan dengan hak-hak asasi lainnya. GDPR menjunjung tinggi hak-hak asasi, termasuk penghormatan terhadap privasi, kebebasan berekspresi, kebebasan hati nurani dan beragama, hak atas perlakuan yang adil, dan penghormatan terhadap keberagaman budaya, agama, dan bahasa.

GDPR adalah kerangka regulasi komprehensif yang mengatur penggunaan data pribadi oleh sektor publik dan swasta, yang mencakup lebih dari sekadar sistem cerdas hingga ke bidang-bidang seperti keamanan nasional, ekonomi, pendidikan, kesehatan, dan budaya. GDPR membedakan penggunaan data pribadi dalam aktivitas pribadi, seperti jejaring sosial, dan konteks profesional atau komersial. Meskipun GDPR berlaku untuk pengendali data, GDPR mengecualikan otoritas yang terlibat dalam investigasi kriminal dan aktivitas keamanan publik.

Penggunaan data pribadi oleh perusahaan atau individu untuk tujuan selain investigasi kriminal dan keamanan publik harus melibatkan persetujuan berdasarkan informasi. GDPR secara luas mendefinisikan persetujuan berdasarkan informasi sebagai tindakan afirmatif yang

jelas di mana seseorang mengizinkan pemrosesan data pribadinya dengan cara yang diberikan secara bebas, spesifik, berdasarkan informasi, dan tidak ambigu. Ini mencakup sarana elektronik maupun pernyataan lisan. Sarana afirmatif meliputi mencentang kotak di situs web, menyetujui ketentuan penggunaan pada platform data, atau metode afirmatif selektif eksplisit lainnya. Penerimaan implisit, kotak yang telah dicentang sebelumnya, atau mekanisme penerimaan pasif lainnya tidak termasuk. Dalam konteks penelitian ilmiah, karena kegunaan data mungkin tidak jelas di awal penelitian, persetujuan berdasarkan informasi mengizinkan penggunaan di area penelitian yang luas, tanpa perlu menentukan secara pasti penggunaan yang diperoleh dari data yang diberikan.

GDPR mengamanatkan bahwa semua pemrosesan data pribadi harus adil, artinya penggunaannya harus transparan kepada individu yang datanya dikumpulkan, digunakan, dikonsultasikan, atau diproses dengan cara lain. Prinsip transparansi mengharuskan semua informasi yang terkait dengan pemrosesan data pribadi mudah diakses dan dijelaskan dalam bahasa yang jelas dan lugas. Prinsip ini terutama menyangkut tujuan pemrosesan data oleh pengendali data, serta informasi yang diperlukan untuk memastikan penggunaan yang wajar dan pemrosesan yang transparan, dengan tetap menghormati hak individu untuk mengakses informasi. GDPR juga menggarisbawahi pentingnya penggunaan data pribadi untuk jangka waktu yang jelas dan terbatas, yang mewajibkan langkah-langkah yang diperlukan untuk menghilangkan atau memperbaiki data pribadi yang tidak akurat.

Prinsip transparansi GDPR menetapkan bahwa informasi yang diakses atau diperoleh dari penggunaan data pribadi harus bersifat publik, mudah diakses, mudah dipahami, dan jika sesuai, menyertakan visualisasi. Informasi tersebut harus dapat diakses secara digital, misalnya melalui situs web. Kerangka regulasi ini mengadvokasi penanganan data olahan yang aman dan transparan terkait individu, dengan mempertimbangkan berbagai keadaan dan konteks pemrosesan data pribadi. Hal ini melibatkan pertimbangan cermat terhadap faktor-faktor yang dapat mengakibatkan ketidakakuratan. Data harus digunakan secara aman dengan cara yang mempertimbangkan potensi risiko terkait kepentingan dan hak subjek data, serta mencegah dampak diskriminatif terhadap individu berdasarkan ras, etnis, pandangan politik, agama, keyakinan, keanggotaan kelompok, status kesehatan, orientasi seksual, atau dampak apa pun yang mungkin timbul dari pembuatan profil.

GDPR memberlakukan kewajiban kepada pengendali di area yang melibatkan penggunaan data pribadi untuk membuat keputusan otomatis di tingkat individu. Individu berhak untuk menolak penggunaan data mereka, terutama ketika keputusan dibuat berdasarkan profil yang dibuat dari data mereka. Pengendali diwajibkan untuk menunjukkan alasan yang sah dan kuat atas penggunaan data pribadi. Individu berhak untuk tidak menjadi subjek keputusan yang semata-mata didasarkan pada pemrosesan otomatis, termasuk pembuatan profil, yang menimbulkan dampak hukum atau dampak signifikan serupa. GDPR mendefinisikan peran petugas perlindungan data, yang ditunjuk oleh pengawas untuk menanggapi permintaan transparansi data, kecuali jika permintaan tersebut bersifat yudisial. Petugas ini bertanggung jawab untuk mengawasi operasi pemantauan pemrosesan data dan memastikan kepatuhan terhadap kerangka peraturan GDPR.

Selain itu, GDPR menguraikan perlunya mekanisme sertifikasi di tingkat negara bagian. Bersamaan dengan itu, GDPR mengamanatkan pembentukan lembaga akreditasi. Lembaga-lembaga ini terlibat dalam aspek akuntabilitas dan transparansi setiap inisiatif terkait penggunaan data pribadi. Tindakan pengawasan dan investigasi, yang didefinisikan sebagai forum hubungan akuntabilitas, disebut audit.

6.2 INISIATIF AWAL DALAM AUDIT AI

Salah satu praktik audit paling awal diinisiasi oleh Kantor Komisioner Informasi (ICO) Inggris untuk memastikan organisasi mematuhi undang-undang perlindungan data, khususnya GDPR. ICO melakukan audit konsensual dan wajib untuk menilai pemrosesan data pribadi dan memberikan panduan tentang perbaikan. Audit ini bertujuan untuk meningkatkan kesadaran akan perlindungan data, menunjukkan komitmen organisasi, dan membangun kepercayaan publik, yang mendorong inovasi dan pertumbuhan.

Inggris mengadopsi GDPR melalui Undang-Undang Perlindungan Data 2018 (DPA 2018), yang mengatur bagaimana informasi pribadi digunakan oleh organisasi, bisnis, dan pemerintah, memastikan bahwa data ditangani secara legal dan etis. Sebagai mitra Inggris untuk GDPR Uni Eropa, DPA 2018 memainkan peran penting dalam mengatur data pribadi di berbagai sektor, melindungi individu dari penggunaan data mereka yang tidak sah atau tidak semestinya. DPA 2018 menetapkan prinsip-prinsip perlindungan data yang ketat yang harus dipatuhi oleh semua entitas yang menangani data pribadi. Prinsip-prinsip ini memastikan data diproses secara adil, sah, transparan, dan untuk tujuan-tujuan tertentu yang telah ditetapkan. Prinsip-prinsip ini juga mensyaratkan pemrosesan data yang memadai, relevan, dan terbatas pada hal-hal yang diperlukan, sekaligus memastikan akurasi data, pembaruan tepat waktu, dan penyimpanan yang aman hanya selama diperlukan. Untuk melindungi data, langkah-langkah keamanan yang tepat harus mencegah pemrosesan, akses, kehilangan, atau kerusakan yang melanggar hukum atau tidak sah

Individu diberikan hak-hak penting berdasarkan DPA 2018, termasuk hak untuk mengakses dan mengoreksi data mereka, meminta penghapusan data dalam kondisi tertentu, dan membatasi atau menolak pemrosesan data. Undang-undang ini juga memberikan perlindungan yang lebih baik untuk data sensitif, seperti asal ras atau etnis, opini politik, dan informasi kesehatan. Perlindungan tambahan diberlakukan untuk data yang terkait dengan hukuman dan pelanggaran pidana, serta hak-hak terkait pengambilan keputusan dan pembuatan profil otomatis. Proses audit ICO dimulai dengan rapat pendahuluan untuk menentukan ruang lingkup dan metodologi audit, yang disesuaikan dengan risiko dan permasalahan spesifik masing-masing organisasi. Ruang lingkup audit biasanya mencakup berbagai area operasional dan manajemen seperti tata kelola, kontrak, pelatihan, dan minimisasi data. Audit ini melibatkan peninjauan dokumen-dokumen relevan, wawancara dengan personel kunci, dan pengamatan langsung terhadap praktik operasional.

Aspek praktis audit meliputi gangguan minimal terhadap operasional harian, dengan ICO memanfaatkan teknik audit jarak jauh jika diperlukan. Aktivitas di lokasi mungkin masih diperlukan untuk inspeksi menyeluruh. Selama audit, tim ICO berinteraksi dengan staf

organisasi untuk memahami dan mengevaluasi implementasi serta efektivitas langkah-langkah perlindungan data.

ICO akan mengidentifikasi dan memprioritaskan pengendali data berisiko tinggi untuk audit berdasarkan berbagai kriteria. Kriteria ini meliputi volume dan sifat pelanggaran yang dilaporkan, pengaduan yang diterima oleh ICO, dan isi laporan tahunan pengendali terkait praktik pengendalian data mereka. Sumber informasi tambahan seperti laporan media dan data publik lainnya juga akan menjadi dasar dalam proses seleksi. Dampak potensial dari ketidakpatuhan dievaluasi dengan mempertimbangkan berapa banyak individu yang terdampak, sensitivitas data yang diproses, dan tingkat kerugian atau tekanan yang mungkin ditimbulkan oleh ketidakpatuhan.

Pendekatan komprehensif ini memungkinkan ICO untuk mengalokasikan sumber dayanya secara efektif kepada organisasi-organisasi dengan risiko kesalahan penanganan data yang paling tinggi. Organisasi yang dipilih untuk diaudit dikategorikan ke dalam lima kelompok: organisasi sukarela, organisasi yang diidentifikasi melalui penilaian risiko, entitas yang menyediakan kesempatan pendidikan yang relevan di bidang-bidang yang menjadi perhatian khusus ICO, organisasi yang direkomendasikan oleh departemen ICO lain untuk diaudit, dan organisasi yang diidentifikasi melalui investigasi oleh Tim Investigasi ICO. Klasifikasi ini memastikan bahwa program audit terarah dan adaptif terhadap risiko yang terus berkembang serta area-area yang menjadi perhatian dalam perlindungan data.

Sepanjang audit, komunikasi rutin dengan staf kunci membantu menilai efektivitas operasional, dilengkapi dengan analisis data dan pengujian kontrol. Jika terjadi pelanggaran data, langkah-langkah segera dikomunikasikan. Pembaruan harian mengenai area yang menjadi perhatian diberikan, yang berpuncak pada rapat penutup untuk membahas isu-isu utama dan langkah-langkah selanjutnya. Akhirnya, draf laporan diterbitkan untuk ditinjau dan disusun rencana aksi oleh organisasi, diikuti dengan penyampaian laporan akhir dan ringkasan eksekutif.

Kelompok Pakar Tingkat Tinggi Kecerdasan Buatan (AI HLEG - 2018)

GDPR berfungsi sebagai kerangka kerja regulasi yang luas untuk melindungi dan mengelola data pribadi, dengan implikasi signifikan bagi AI. Mengingat pentingnya AI, Parlemen Eropa menugaskan sekelompok pakar untuk mengembangkan Strategi AI untuk Eropa. Inisiatif ini, yang melibatkan anggota dari industri, akademisi, dan masyarakat sipil, bertujuan untuk memastikan keberagaman, koherensi, dan konsistensi dalam pendekatan Eropa terhadap AI. Kelompok tersebut, yang dikenal sebagai AI High-Level Expert Group (AI HLEG), merilis pedoman pada Desember 2018 untuk AI yang Tepercaya. Menurut dokumen tersebut, AI yang Tepercaya harus mematuhi hukum dan peraturan yang berlaku, menjunjung tinggi prinsip dan nilai etika, serta menunjukkan ketahanan baik secara teknis maupun dampak sosialnya.

Pada April 2019, kelompok tersebut menerbitkan pedoman etika untuk AI yang Tepercaya. Dokumen tersebut menguraikan kerangka kerja dengan mendefinisikan fundamentalnya dan kemudian merinci rencana aksi untuk mencapainya. Landasan tersebut menekankan pentingnya mengembangkan, menerapkan, dan memanfaatkan sistem AI yang

menghormati prinsip-prinsip etika seperti otonomi manusia, pencegahan bahaya, keadilan, dan keterjelasan. Pedoman ini juga menyoroti perlunya mengenali dan mengatasi potensi ketegangan di antara prinsip-prinsip ini dan memberikan perhatian khusus pada skenario yang memengaruhi kelompok rentan, seperti anak-anak, penyandang disabilitas, dan mereka yang secara historis kurang beruntung atau berisiko dikucilkan.

Lebih lanjut, pedoman tersebut menekankan perlunya mengakui adanya asimetri kekuasaan atau informasi, seperti antara pemberi kerja dan pekerja atau antara bisnis dan konsumen. Dokumen tersebut juga mencatat bahwa meskipun sistem AI dapat menawarkan manfaat substansial bagi individu dan masyarakat, sistem ini juga menimbulkan risiko tertentu dan dapat menimbulkan dampak negatif, yang beberapa di antaranya mungkin sulit diprediksi, diidentifikasi, atau diukur. Misalnya, dampaknya terhadap demokrasi, supremasi hukum, keadilan distributif, atau pikiran manusia itu sendiri. Oleh karena itu, langkah-langkah yang tepat harus diambil untuk memitigasi risiko-risiko ini, dengan tingkat keparahan langkah-langkah yang proporsional dengan tingkat risikonya.

Rencana aksi dalam dokumen tersebut menguraikan tujuh persyaratan yang harus dipenuhi oleh sistem AI untuk mencapai AI Tepercaya. Pedoman ini mencakup memastikan bahwa pengembangan, penerapan, dan penggunaan sistem AI mematuhi standar-standar AI yang andal, yaitu: (1) peran serta manusia dan pengawasan, (2) ketahanan dan keamanan teknis, (3) privasi dan tata kelola data, (4) transparansi, (5) keberagaman, non-diskriminasi, dan keadilan, (6) kesejahteraan masyarakat dan lingkungan, serta (7) akuntabilitas. Untuk memastikan persyaratan ini terpenuhi, perlu dipertimbangkan penggunaan metode teknis dan non-teknis.

Perlu juga ada dorongan untuk meningkatkan penelitian dan inovasi guna menilai sistem AI dan mendorong kepatuhan; hasil dan pertanyaan interpretasi terbuka harus diungkapkan kepada publik, dan generasi baru spesialis Etika AI harus dilatih secara sistematis. Kebutuhan untuk mengomunikasikan informasi secara jelas dan proaktif kepada para pemangku kepentingan tentang kapabilitas dan keterbatasan sistem AI ditekankan, yang memfasilitasi penetapan ekspektasi yang realistis dan pemahaman tentang kepatuhan terhadap persyaratan. Transparansi tentang penggunaan sistem AI, memastikan ketertelusuran dan auditabilitas, terutama dalam konteks atau situasi kritis, sangatlah penting. Pelibatan para pemangku kepentingan di seluruh siklus hidup sistem AI juga diperlukan, dengan mendorong pendidikan dan pelatihan agar semua pihak memahami Trustworthy AI dan menerima instruksi yang tepat. Terakhir, penting untuk mengakui bahwa mungkin terdapat ketegangan mendasar antara berbagai prinsip dan persyaratan; ketegangan ini dan penyelesaiannya harus diidentifikasi, dievaluasi, didokumentasikan, dan dikomunikasikan secara konsisten.

Dokumen ketiga yang diterbitkan oleh kelompok tersebut pada Juli 2020 membahas persyaratan yang diperlukan untuk mencapai Trustworthy AI. Persyaratan ini dirinci dalam "Glossary of Assessment List for Trustworthy Artificial Intelligence (ALTAI)" dan pada dasarnya memperluas persyaratan yang diuraikan dalam dokumen pedoman etika untuk Trustworthy AI. Persyaratan ini meliputi:

- Agensi dan Pengawasan Manusia: Sistem AI harus mendukung tindakan dan pengambilan keputusan manusia, dengan berpegang teguh pada prinsip menghormati otonomi manusia. Hal ini mengharuskan sistem AI bertindak sebagai pendorong bagi masyarakat yang demokratis, sejahtera, dan berkeadilan, serta menjunjung tinggi hak-hak dasar yang didukung oleh pengawasan manusia.
- **Ketahanan dan Keamanan Teknis**: Persyaratan penting untuk mencapai sistem AI yang andal adalah kepercayaan (kemampuan untuk menyediakan layanan yang dapat dipercaya) dan ketahanan (ketahanan dalam menghadapi perubahan). Ketahanan teknis menuntut sistem AI dikembangkan dengan pendekatan preventif terhadap risiko dan berfungsi dengan andal sesuai rencana, sekaligus meminimalkan kerugian yang tidak diinginkan dan tak terduga serta mencegahnya jika memungkinkan. Hal ini terutama berlaku dalam skenario yang melibatkan potensi perubahan dalam lingkungan operasionalnya atau interaksi dengan agen lain (manusia atau buatan) yang dapat berdampak buruk pada sistem AI. Persyaratan ini mencakup empat aspek:
    1. **Keamanan**: Ini menyangkut kemungkinan sistem cerdas menyebabkan kerugian bagi manusia.
    2. **Keselamatan**: Ini berkaitan dengan penggunaan metrik dan penilaian tingkat risiko selama pengembangan sistem cerdas.
    3. **Akurasi**: Ini melibatkan hubungan antara kerugian dan tingkat akurasi yang rendah dalam sistem.
    4. **Keandalan, rencana cadangan, dan reproduktifitas**: Ini berkaitan dengan risiko dan kerusakan yang disebabkan oleh sistem akibat hasil yang tidak andal atau tidak dapat direproduksi.
- **Privasi dan Tata Kelola Data**: Privasi, hak fundamental yang khususnya terdampak oleh sistem AI, sangat erat kaitannya dengan pencegahan kerugian. Mencegah kerugian terhadap privasi memerlukan tata kelola data yang tepat, yang mencakup kualitas dan integritas data yang digunakan, relevansinya mengingat domain implementasi sistem AI, protokol aksesnya, dan kemampuan untuk memprosesnya dengan cara yang melindungi privasi. Persyaratan ini terkait dengan GDPR.
- **Transparansi**: Komponen krusial untuk mencapai AI Tepercaya adalah transparansi, yang mencakup tiga elemen:
    1. **Ketertelusuran**: Ini melibatkan penilaian apakah proses pengembangan sistem AI, yaitu data dan proses yang menghasilkan keputusan sistem, didokumentasikan secara memadai untuk memungkinkan ketertelusuran, meningkatkan transparansi, dan pada akhirnya membangun kepercayaan terhadap AI di masyarakat.
    2. **Keterjelasan**: Ini adalah kemampuan untuk menjelaskan proses teknis sistem AI dan alasan di balik keputusan atau prediksi yang dibuatnya. Keterjelasan sangat penting untuk membangun dan menjaga kepercayaan pengguna terhadap sistem AI. Keputusan yang digerakkan oleh AI harus dijelaskan dan

dipahami, sebisa mungkin, oleh mereka yang secara langsung maupun tidak langsung terdampak agar keputusan tersebut dapat ditantang.

3. **Komunikasi tentang keterbatasan**: Ini melibatkan penilaian apakah kapabilitas dan keterbatasan sistem AI telah dikomunikasikan kepada pengguna dengan tepat untuk kasus penggunaan yang sedang dihadapi. Ini dapat mencakup pengomunikasian tingkat akurasi sistem AI, serta keterbatasannya.

- **Keberagaman, Nondiskriminasi, dan Keadilan**: Untuk mencapai AI yang Tepercaya, kita harus mendorong inklusi dan keberagaman di seluruh siklus hidup sistem AI. Sistem AI, baik selama pelatihan maupun pengoperasian, dapat mengandung bias historis yang tidak disengaja, model yang tidak lengkap, dan tata kelola yang buruk. Bias yang berkelanjutan dapat menyebabkan prasangka dan diskriminasi langsung terhadap kelompok atau individu tertentu, yang berpotensi memperburuk bias dan marginalisasi. Kerugian juga dapat timbul dari eksploitasi bias yang disengaja atau terlibat dalam persaingan tidak sehat, seperti homogenisasi harga melalui kolusi atau pasar yang tidak transparan. Sedapat mungkin, bias yang dapat diidentifikasi dan diskriminatif harus dihilangkan selama fase pengumpulan data. Sistem AI harus berpusat pada pengguna dan dirancang untuk memungkinkan semua individu menggunakan produk atau layanan AI, tanpa memandang usia, jenis kelamin, kemampuan, atau karakteristik. Memastikan aksesibilitas bagi penyandang disabilitas, yang hadir di semua kelompok sosial, sangatlah penting. Aspek ini terkait erat dengan bias data dan bias algoritmik, yang keduanya melibatkan bagaimana sistem cerdas berbasis data dapat mengumpulkan data yang melanggengkan bias, sehingga mereproduksi stereotip dan memperlebar kesenjangan dalam perlakuan yang tidak setara yang memengaruhi individu atau kelompok dalam masyarakat. Selain itu, sistem AI tidak boleh mengadopsi pendekatan satu ukuran untuk semua dan harus merangkul prinsip-prinsip Desain Universal, menangani rentang pengguna terluas dan mematuhi standar aksesibilitas yang relevan. Ini akan memungkinkan akses yang adil dan partisipasi aktif semua orang dalam sistem komputasi yang ada dan yang sedang berkembang.
- **Kesejahteraan Masyarakat dan Lingkungan**: Sejalan dengan prinsip-prinsip keadilan dan pencegahan bahaya, masyarakat luas, makhluk hidup berakal lainnya, dan lingkungan harus dianggap sebagai pemangku kepentingan di sepanjang siklus hidup sistem AI. Paparan yang meluas terhadap sistem AI sosial di semua bidang kehidupan kita dapat berdampak negatif pada hubungan dan keterikatan sosial kita. Meskipun sistem AI dapat meningkatkan keterampilan sosial, sistem ini juga dapat berkontribusi pada kemerosotannya. Hal ini dapat memengaruhi kesejahteraan fisik dan mental individu. Oleh karena itu, pemantauan dan pertimbangan cermat terhadap dampak sistem AI sangatlah penting. Selain itu, keberlanjutan dan tanggung jawab ekologis harus dipromosikan, dan penelitian tentang solusi AI yang menangani area-area yang menjadi perhatian global harus didorong. Secara umum, AI harus digunakan untuk memberi manfaat bagi seluruh umat manusia, termasuk generasi mendatang. Sistem

AI harus berfungsi untuk memelihara dan mempromosikan proses demokrasi serta menghormati pluralitas nilai dan pilihan hidup individu. Sistem AI tidak boleh merusak proses demokrasi, musyawarah manusia, atau sistem pemungutan suara demokratis, atau menimbulkan ancaman sistemik bagi masyarakat luas. Poin terakhir ini mendapat penekanan khusus. Penting untuk menilai sendiri dampak sistem AI dari perspektif sosial, dengan mempertimbangkan dampaknya terhadap institusi, demokrasi, dan masyarakat luas, misalnya, ketika sistem AI memperburuk berita palsu, memisahkan pemilih, atau memfasilitasi perilaku totaliter.

- **Akuntabilitas**: Prinsip akuntabilitas mengamanatkan pembentukan mekanisme untuk memastikan tanggung jawab atas pengembangan, penerapan, dan/atau penggunaan sistem AI. Isu ini berkaitan erat dengan manajemen risiko, yang melibatkan identifikasi dan mitigasi risiko secara transparan yang dapat dijelaskan dan diaudit oleh pihak ketiga. Jika terjadi dampak yang tidak adil atau merugikan, perlu ada mekanisme akuntabilitas yang dapat diakses untuk memastikan kesempatan yang tepat untuk perbaikan. Persyaratan ini menekankan auditabilitas AI, yang melibatkan penilaian mandiri atas tingkat saat ini atau yang diperlukan untuk mengevaluasi sistem AI oleh auditor internal dan eksternal. Kemampuan untuk melakukan evaluasi dan akses ke evaluasi ini dapat berkontribusi pada AI yang Tepercaya. Dalam aplikasi yang memengaruhi hak-hak dasar, termasuk yang penting bagi keselamatan, sistem AI harus dirancang untuk memungkinkan audit independen.

Dokumen tersebut menguraikan beberapa konsep dasar. Misalnya, dokumen tersebut mendefinisikan akuntabilitas sebagai "gagasan bahwa seseorang bertanggung jawab atas tindakannya. dan, akibatnya, hasilnya, dan harus mampu menjelaskan tujuan, motivasi, dan alasan mereka." Akuntabilitas memiliki beragam dimensi dan terkadang diamanatkan oleh hukum. Misalnya, GDPR mewajibkan organisasi yang menangani data pribadi untuk menerapkan langkah-langkah keamanan guna mencegah pelanggaran data dan mewajibkan pelaporan jika langkah-langkah tersebut gagal. Namun, akuntabilitas juga dapat mencerminkan standar etika yang belum tentu berujung pada konsekuensi hukum. Perusahaan teknologi tertentu mungkin memilih untuk tidak mengembangkan teknologi pengenalan wajah, meskipun tidak ada larangan atau moratorium teknologi, berdasarkan pertimbangan etika akuntabilitas.

Dokumen tersebut juga mendefinisikan keadilan sebagai sesuatu yang mencakup berbagai konsep yang dikenal sebagai kesetaraan, imparsialitas, egalitarianisme, non-diskriminasi, dan keadilan. Keadilan terutama berkaitan dengan cita-cita perlakuan yang setara di antara individu atau kelompok, yang sering disebut sebagai keadilan 'substantif'. Selain itu, keadilan mencakup aspek prosedural, yang melibatkan kemampuan untuk mencari dan memperoleh ganti rugi ketika hak dan kebebasan individu dilanggar.

Kerangka regulasi Uni Eropa tentang AI (Undang-Undang AI Uni Eropa)

Upaya selanjutnya oleh Parlemen Eropa bertujuan untuk merancang strategi regulasi guna menetapkan batasan dan tanggung jawab dalam penerapan AI Tepercaya. Pada tanggal 21 April 2021, dokumen “Menetapkan aturan yang diharmonisasikan tentang AI dan

mengubah undang-undang serikat tertentu" telah dirilis. Dokumen ini mencakup memorandum penjelasan yang menyertai peraturan yang diusulkan. Memorandum tersebut menguraikan faktor-faktor kontekstual yang membenarkan promosi proposal tersebut.

Dinyatakan bahwa kebutuhan muncul karena AI dapat mendukung hasil yang bermanfaat secara sosial dan lingkungan serta memberikan keunggulan kompetitif bagi perusahaan, sehingga menguntungkan ekonomi Eropa. Undang-undang ini berdampak pada sektor-sektor penting seperti perubahan iklim, kesehatan, lingkungan, sektor publik, dan pertanian. Terlepas dari manfaatnya, diakui bahwa AI menimbulkan risiko, yang memerlukan pendekatan yang seimbang terhadap regulasi AI. Kebutuhan untuk mendefinisikan regulasi bermula dari buku putih Komisi Eropa "Sebuah persatuan yang berjuang untuk lebih", yang menyoroti perlunya legislasi untuk pendekatan terkoordinasi terhadap implikasi etis AI. Menyusul pengumuman ini, pada 19 Februari 2020, Komisi Eropa menerbitkan buku putih "Pendekatan Eropa terhadap Keunggulan dan Kepercayaan". Buku putih ini membahas upaya mempromosikan pengembangan AI sekaligus mempertimbangkan risikonya.

Konteks tambahan, seperti kesimpulan dari Dewan Eropa pada 21 Oktober 2020, menyoroti perlunya mengatasi ketidakjelasan, kompleksitas, bias, dan tingkat otonomi sistem AI tertentu untuk memastikan keselarasan dengan hak-hak fundamental dan kepatuhan terhadap kerangka hukum. Dewan Eropa juga memberikan latar belakang untuk proposal tersebut berdasarkan resolusi terkait AI, yang membahas isu-isu terkait etika, hak cipta, dan hal-hal terkait privasi lainnya dalam kerangka GDPR. Berdasarkan semua konteks ini, Komisi Eropa mengusulkan kerangka kerja regulasi AI dengan tujuan-tujuan berikut:

- Memastikan keamanan sistem AI dan mematuhi kerangka hukum yang berlaku,
- Menjamin kepastian hukum untuk memfasilitasi investasi dan inovasi di bidang AI,
- Meningkatkan tata kelola dan penegakan hukum yang efektif berdasarkan hak-hak fundamental yang berlaku untuk sistem AI,
- Memfasilitasi pengembangan aplikasi AI yang tepercaya untuk mencegah fragmentasi pasar.
- Memorandum tersebut menetapkan prinsip utama untuk perancangan proposal regulasi:

> *Prinsip utama untuk perancangan proposal regulasi*
> Upaya regulasi harus dibatasi pada persyaratan minimum yang diperlukan untuk mengatasi risiko yang terkait dengan AI tanpa menghambat kemajuan teknologi.

Tantangannya adalah menciptakan proposal berwawasan ke depan yang tetap relevan dengan mengantisipasi tantangan teknologi di masa depan. Selain itu, proposal tersebut harus secara jelas mendefinisikan kerangka kerja regulasi yang seimbang dan proporsional yang berfokus pada risiko yang terdefinisi dengan baik untuk menghindari pembatasan yang tidak perlu. Intinya, kerangka kerja tersebut harus tepat dengan batasan yang jelas dan mekanisme yang fleksibel yang tidak menghambat pengembangan AI.

Proposal ini menetapkan aturan untuk pengembangan, pemosisian pasar, dan penggunaan sistem AI berdasarkan pendekatan berbasis risiko. Proposal ini melarang praktik-praktik berbahaya tertentu dan menetapkan beberapa pengecualian, termasuk penggunaan khusus sistem identifikasi biometrik jarak jauh untuk penegakan hukum. Aspek kunci dari proposal ini adalah mendefinisikan apa yang dimaksud dengan sistem AI berisiko tinggi. Sistem tersebut harus mematuhi standar yang telah ditetapkan dalam proposal agar dapat beroperasi. Berbagai kewajiban juga dibebankan kepada penyedia dan pengguna sistem ini untuk memastikan keamanan penggunaannya. Untuk beberapa sistem AI tertentu, hanya kewajiban transparansi minimal yang ditetapkan, terutama yang melibatkan chatbot dan deep fake.

Elemen fundamental lain dari proposal ini adalah definisi skema tata kelola. Tata kelola untuk kerangka regulasi ini menetapkan partisipasi negara-negara anggota Uni Eropa, yang menguraikan mekanisme kerja sama melalui Dewan AI Eropa. Dasar hukum proposal ini bertumpu pada kebutuhan untuk mendorong fungsi pasar internal, yang, karena besarnya tantangan, tidak dapat ditangani secara efektif oleh masing-masing negara secara independen. Tujuan proposal ini ditetapkan di tingkat Uni Eropa untuk mencegah proliferasi peraturan daerah yang saling bertentangan, yang akan menciptakan ketidakpastian hukum dan menghambat investasi dalam pengembangan AI.

Mencerminkan semua upaya yang dipimpin oleh Uni Eropa, sebagaimana ditunjukkan dalam inisiatif HLEG dan ALTAI, Aliansi AI dibentuk, yang terdiri dari 4.000 pemangku kepentingan. Kelompok ini berkumpul untuk membahas implikasi sosial dan teknologi AI, yang berpuncak pada pertemuan tahunan. Proposal ini dianggap partisipatif, menarik, dan inklusif, serta melibatkan para pemangku kepentingan dari berbagai sektor yang terkait dengan AI. Berdasarkan musyawarah, diputuskan untuk mengembangkan instrumen legislatif Uni Eropa Horizontal yang mengikuti pendekatan berbasis risiko yang proporsional, beserta kode etik untuk sistem AI non-risiko tinggi. Aspek utama proposal ini adalah menguraikan pendekatan berbasis risiko dan menetapkan pedoman umum untuk pengembangan dan penggunaan sistem non-risiko tinggi, sehingga menciptakan kerangka kerja yang komprehensif untuk operasional AI.

Elemen Dasar Undang-Undang AI Uni Eropa

Proposal ini menguraikan daftar praktik AI yang dilarang. Praktik-praktik ini dijelaskan dalam Judul II, Pasal 5:

1. Pemasaran, baik melalui penawaran layanan maupun penggunaan sistem AI, yang didasarkan pada teknik-teknik subliminal yang bertujuan mendistorsi perilaku seseorang hingga merugikan mereka, baik melalui kerugian materiil maupun psikologis.
2. Pemasaran, baik melalui penawaran layanan maupun penggunaan sistem AI, yang mengeksploitasi kerentanan kelompok tertentu berdasarkan usia, disabilitas fisik, maupun mental, dengan cara yang mendistorsi perilaku mereka dan menyebabkan kerugian materiil maupun psikologis.

3. Pemasaran, baik melalui penawaran layanan maupun penggunaan sistem AI, yang bergantung pada keputusan di ranah publik yang mengevaluasi atau mengklasifikasikan keandalan seseorang dengan skor sosial, yang dapat mengakibatkan perlakuan yang merugikan atau tidak menguntungkan terhadap individu atau kelompok dalam konteks yang tidak terkait dengan pengumpulan data awal atau menyebabkan perlakuan yang tidak proporsional atau tidak adil berdasarkan perilaku sosial mereka.
4. Penggunaan sistem identifikasi biometrik jarak jauh secara real-time di ruang publik, kecuali benar-benar diperlukan. Penggunaan yang diperlukan didefinisikan sebagai penargetan seseorang yang berpotensi menjadi korban kejahatan, termasuk anak hilang; mencegah ancaman terhadap integritas pribadi, seperti dari tindakan teroris; dan pendeteksian, lokasi, identifikasi, dan penuntutan pelaku atau tersangka tindak pidana yang dapat dihukum.

Daftar larangannya singkat namun signifikan. Dua larangan pertama membahas asimetri yang terjadi antara sistem dan penggunanya, baik karena penggunaan strategi subliminal maupun ketidakseimbangan antara kemampuan sistem AI dalam menghasilkan konten dan kapasitas pengguna untuk memproses konten tersebut, misalnya, karena disabilitas atau kerentanan. Kami akan mengeksplorasi isu-isu ini lebih lanjut di bab-bab mendatang, karena isu-isu ini fundamental bagi apa yang kita kenal sebagai disinformasi. Fokus kami adalah mengungkap hubungan antara disinformasi dan AI, serta bagaimana sistem AI dapat memperburuk fenomena ini.

Larangan ketiga membahas penggunaan data di berbagai konteks. Dalam AI, hal ini agak terkait dengan transfer learning, yang melibatkan penerapan model AI di domain yang berbeda dari domain tempat awalnya dilatih. Meskipun larangan ini tidak menargetkan penggunaan data dalam konteks yang berbeda, larangan ini secara khusus membahas pembentukan skor sosial yang mendukung pengambilan keputusan berdasarkan kepercayaan sosial. Dapat dipahami bahwa kasus aplikasi dapat melibatkan penggunaan data dari, misalnya, jejaring sosial untuk menyimpulkan penyakit seseorang.

Informasi ini tidak boleh digunakan dalam konteks yang berbeda, seperti menganalisis akses ke asuransi kesehatan. Inferensi lintas domain untuk mendukung pengambilan keputusan, terutama di ranah publik, merupakan penggunaan data yang kontroversial yang bahkan berdampak pada privasi pribadi. Larangan keempat pada dasarnya menetapkan kerangka kerja untuk penggunaan sistem biometrik waktu nyata (real-time), berdasarkan konsep penggunaan yang diperlukan (necessary use). Aspek kunci lain dari proposal ini melibatkan pengklasifikasian sistem AI berisiko tinggi. Definisi-definisi ini, yang tercantum dalam Judul III proposal, mengkarakterisasi sistem AI berisiko tinggi berdasarkan area penggunaannya. Area-area ini meliputi:

- Identifikasi dan klasifikasi biometrik orang perseorangan.
- Manajemen dan pengoperasian infrastruktur kritis.
- Pendidikan dan pelatihan vokasional.
- Ketenagakerjaan, manajemen sumber daya manusia, dan akses ke wirausaha.

- Sistem seleksi personel, evaluasi kredit, dan prioritas untuk tanggap darurat.
- Sistem yang menyediakan dukungan bagi sistem peradilan.
- Sistem kontrol perbatasan dan migrasi.
- Sistem pendukung bagi proses demokrasi.

Proposal ini membahas definisi tugas dan persyaratan yang harus dipenuhi oleh sistem AI berisiko tinggi. Proposal ini menetapkan bahwa sistem harus diimplementasikan, didokumentasikan, dan dipelihara dengan mempertimbangkan status risiko tingginya. Dalam konteks ini, proposal ini menetapkan bahwa sistem AI berisiko tinggi harus menggabungkan proses yang berkelanjutan dan berulang di seluruh siklus hidupnya, yang mencakup:

1. Identifikasi dan analisis risiko yang diketahui dan potensial yang terkait dengan sistem;
2. Estimasi dan evaluasi risiko yang mungkin muncul ketika sistem AI berisiko tinggi digunakan, baik dalam kondisi yang diinginkan maupun skenario lainnya;
3. Penilaian risiko yang muncul berdasarkan analisis data pemantauan pasca-pemasaran, dan
4. Penerapan langkah-langkah mitigasi risiko.

Persyaratan ini menetapkan sistem pemantauan pasca-pemasaran yang harus secara sistematis mengumpulkan informasi dari pengguna tentang kinerja sistem di seluruh siklus hidupnya. Setelah risiko didefinisikan, penting untuk memastikan penghapusan atau pengurangan risiko baik dalam desain maupun implementasi sistem.

Bila diperlukan, langkah-langkah pengendalian dan mitigasi harus diterapkan untuk risiko yang tidak dapat dihilangkan. Selain itu, pengguna harus diinformasikan tentang risiko-risiko ini. Terkait poin terakhir ini, informasi yang diberikan kepada pengguna harus mendukung operasional sistem yang transparan sehingga mereka dapat menginterpretasikan hasil sistem dan menggunakannya dengan tepat. Ini mencakup petunjuk penggunaan, tujuan sistem, tingkat akurasi, ketahanan, dan keamanan sibernya, kinerjanya terkait kelompok spesifik yang ditargetkan, dan, jika berlaku, spesifikasi terkait data yang digunakan untuk pelatihan, validasi, dan pengujian.

Aspek krusial lain dari proposal ini menyangkut data dan tata kelola data. Telah ditetapkan bahwa partisi data untuk pelatihan, validasi, dan pengujian harus tunduk pada praktik tata kelola dan manajemen data yang tepat. Praktik-praktik ini harus mencakup berbagai tahapan siklus hidup data, termasuk pengumpulan, anotasi data, pelabelan, pembersihan, pengayaan, dan agregasi, tinjauan yang ketat berdasarkan potensi bias dan identifikasi kemungkinan kesenjangan atau kekurangan data, serta bagaimana kesenjangan ini dapat diatasi.

Proposal ini juga mendefinisikan pentingnya pengawasan manusia. Pengawasan ini harus mencegah atau meminimalkan risiko terhadap kesehatan, keselamatan, atau hak-hak asasi yang timbul ketika sistem AI berisiko tinggi digunakan. Pengawasan manusia harus dipastikan sebelum sistem dirilis ke pasar.

Proposal ini menguraikan semua kewajiban yang diemban oleh penyedia dan semua pihak yang terlibat dalam siklus hidup sistem AI berisiko tinggi. Pihak-pihak ini meliputi otoritas yang berwenang, produsen produk, importir, distributor, dan pengguna. Setiap pelaku

memiliki serangkaian kewajiban yang ditetapkan dalam proposal. Untuk memastikan kepatuhan terhadap persyaratan dan tugas yang ditetapkan, proposal ini mengacu pada standar dan penilaian kesesuaian. Proposal ini menekankan perlunya mendefinisikan standar yang diselaraskan, yang harus selaras dengan persyaratan yang ditentukan dalam proposal. Jika penyedia mengadopsi standar yang diselaraskan ini, mereka harus mengikuti prosedur penilaian kesesuaian yang didasarkan pada strategi pengendalian internal atau sistem jaminan mutu. Sistem jaminan mutu harus mencakup elemen-elemen berikut:

1. Strategi untuk kepatuhan regulasi, termasuk prosedur kesesuaian.
2. Teknik, prosedur, dan tindakan sistematis yang akan digunakan untuk perancangan dan verifikasi sistem AI berisiko tinggi.
3. Teknik, prosedur, dan tindakan sistematis yang akan digunakan dalam pengembangan, pengendalian mutu, dan jaminan mutu sistem AI berisiko tinggi. 4. Pemeriksaan, pengujian, dan validasi yang akan dilakukan sebelum, selama, dan setelah pengembangan sistem.
4. Spesifikasi teknis, termasuk standar yang akan diterapkan dan metode untuk memastikan kepatuhan terhadap tujuan proposal ketika standar yang diharmonisasikan tidak sepenuhnya diterapkan.
5. Sistem dan prosedur untuk manajemen data, yang mencakup semua tahapan siklus hidup data.
6. Sistem manajemen risiko.
7. Implementasi dan pemeliharaan sistem setelah memasuki pasar.
8. Prosedur terkait pelaporan insiden serius.
9. Manajemen komunikasi dengan otoritas yang berwenang.
10. Sistem dan prosedur untuk mendokumentasikan semua informasi yang relevan.
11. Manajemen sumber daya, termasuk langkah-langkah terkait keamanan pasokan.
12. Kerangka kerja akuntabilitas yang menguraikan tanggung jawab manajemen dan staf lainnya.

Proposal ini juga menetapkan kewajiban transparansi untuk sistem AI lainnya (Judul IV). Khususnya, undang-undang ini mewajibkan penyedia layanan untuk memastikan bahwa sistem AI yang berinteraksi dengan individu dirancang sedemikian rupa sehingga individu tersebut mendapatkan informasi saat berinteraksi dengan sistem AI, kecuali jika hal tersebut jelas dari konteks penggunaannya. Kewajiban ini tidak berlaku untuk sistem yang secara hukum berwenang untuk mendeteksi dan mencegah aktivitas kriminal. Kewajiban lain yang terkait dengan transparansi mewajibkan pengguna sistem pengenalan emosi atau sistem kategorisasi biometrik untuk memberi tahu individu bahwa mereka tunduk pada sistem tersebut.

Hal ini tidak berlaku untuk sistem kategorisasi biometrik tertentu yang secara hukum berwenang untuk mendeteksi dan mencegah aktivitas kriminal. Lebih lanjut, pengguna sistem AI yang menghasilkan atau memanipulasi gambar, audio, atau video yang menyerupai orang, objek, tempat, atau entitas lain dengan cara yang dapat menyesatkan dan tampak sebagai konten asli ('deep fake') harus mengungkapkan bahwa konten tersebut telah dihasilkan atau

dimanipulasi secara artifisial. Terakhir, proposal ini menetapkan tata kelola regulasi, yang dipimpin oleh Dewan Kecerdasan Buatan Eropa, yang bertugas membantu dan memberi nasihat kepada Komisi Eropa mengenai aspek kerja sama dan koordinasi yang diperlukan untuk menerapkan kerangka regulasi. Peraturan ini juga menetapkan Otoritas Kompeten Nasional, yang harus menjaga kerahasiaan informasi dan data yang diperoleh selama penegakan peraturan.

Pengenaan denda dan sanksi administratif berada di tangan negara-negara anggota Uni Eropa. Sanksi harus efektif, proporsional, dan bersifat jera. Sanksi harus mempertimbangkan kepentingan penyedia skala kecil dan perusahaan rintisan dengan cara yang tidak mengorbankan kelangsungan ekonomi mereka. Sanksi ekonomi dapat mencapai hingga 30 juta euro atau, jika pelanggarnya adalah perusahaan, hingga 6% dari total omzet tahunan globalnya untuk tahun keuangan sebelumnya, mana yang lebih tinggi, jika terdapat ketidakpatuhan terhadap larangan praktik AI atau perlindungan data dan tata kelola data.

Ketidakpatuhan terhadap kewajiban lain yang ditetapkan dalam peraturan akan dikenakan sanksi ekonomi hingga 20 juta euro atau, jika pelanggarnya adalah perusahaan, hingga 4% dari total omzet tahunan globalnya untuk tahun keuangan sebelumnya. Informasi yang menyesatkan juga dapat dikenakan denda hingga 10 juta euro atau, jika pelakunya adalah perusahaan, hingga 2% dari total omzet tahunan globalnya pada tahun keuangan sebelumnya. Pada saat buku ini ditulis, versi final Undang-Undang AI telah diterbitkan pada Desember 2023 dan mencakup sebagian besar aspek yang awalnya diuraikan dalam diskusi ekstensif tentang prinsip-prinsip dasar undang-undang tersebut, yang dirangkum dalam bagian ini. Diskusi luas ini, yang menghasilkan buku putih tentang Undang-Undang AI, melewati fase konsolidasi, yang menghasilkan Undang-Undang AI Uni Eropa. Undang-undang ini merupakan salah satu kerangka regulasi transnasional pertama dalam AI.

6.3 AUDIT DAN TATA KELOLA AI

Meningkatnya risiko yang terkait dengan AI telah mendorong berbagai lembaga untuk mengembangkan kerangka kerja yang mematuhi prinsip-prinsip AI yang bertanggung jawab. Kebutuhan untuk menunjukkan dan mensertifikasi kepatuhan terhadap prinsip-prinsip AI yang bertanggung jawab telah mendorong pengembangan inisiatif audit AI. Audit ini menawarkan berbagai metode untuk memeriksa sistem sosioteknis AI yang mungkin memiliki dampak sosial. Audit AI harus mendokumentasikan dampak-dampak ini dan bagaimana dampak tersebut terjadi.

Audit AI terutama berfokus pada verifikasi kepatuhan pengembang terhadap prinsip-prinsip transparansi, keadilan, dan akuntabilitas di seluruh siklus hidup AI, serta kepatuhan terhadap undang-undang, peraturan, dan pedoman lain yang berasal dari kebijakan publik. Sebagaimana diilustrasikan oleh proposal peraturan Uni Eropa, yang mendefinisikan larangan dan daftar persyaratan untuk pengembangan dan penggunaan sistem berisiko tinggi, verifikasi kepatuhan terhadap persyaratan semacam ini menjadi semakin penting. Inisiatif audit AI bertujuan untuk memverifikasi kepatuhan proses dan hasil dalam kerangka peraturan yang ada.

Audit AI didukung di bawah payung tata kelola AI. Berbagai lembaga tata kelola AI menyediakan dokumen berisi rekomendasi atau standar yang harus dipatuhi oleh pengembang. Hal serupa juga terjadi pada Parlemen Uni Eropa dan Undang-Undang AI Uni Eropa. Bersamaan dengan inisiatif ini, lembaga-lembaga lain telah bergabung dalam upaya ini.

Kantor Akuntabilitas Pemerintah (GAO) yang berbasis di AS

GAO mengembangkan kerangka kerja akuntabilitas bagi lembaga federal dan entitas lain untuk memastikan akuntabilitas dan penggunaan AI yang bertanggung jawab dalam program dan proses pemerintah. Kerangka kerja ini disusun berdasarkan empat prinsip yang membahas tata kelola, data, kinerja, dan pemantauan sistem AI. Untuk setiap prinsip, kerangka kerja ini menguraikan praktik-praktik utama bagi lembaga federal, lembaga lain, dan pengembang sistem AI. Kerangka kerja ini mencakup praktik-praktik penting, pertanyaan untuk pengembang, dan prosedur untuk auditor.

Tata Kelola: Kerangka kerja ini mendefinisikan tata kelola sebagai pengelolaan dan pengawasan AI oleh mereka yang bertanggung jawab, yang dapat memanfaatkan struktur tata kelola untuk mengelola risiko, menunjukkan integritas dan kepatuhan terhadap nilai-nilai etika, serta memastikan kepatuhan terhadap hukum dan peraturan yang relevan. Kerangka kerja ini mengidentifikasi sembilan praktik yang dikelompokkan di tingkat organisasi dan sistem. Di tingkat organisasi, manajer diharapkan untuk membangun lingkungan yang menumbuhkan sikap positif dan proaktif terhadap pengendalian internal. Terdapat enam praktik utama yang membantu menetapkan prinsip-prinsip ini di tingkat organisasi:

1. **Tujuan yang jelas**: Tetapkan tujuan yang jelas untuk sistem AI guna memastikan tercapainya hasil yang diharapkan.
2. **Peran dan tanggung jawab:** Tetapkan peran dan tanggung jawab yang jelas untuk memastikan operasi yang efektif, koreksi yang tepat waktu, dan pengawasan.
3. **Nilai**: Tunjukkan komitmen terhadap nilai dan prinsip yang ditetapkan oleh entitas untuk meningkatkan kepercayaan publik terhadap penggunaan AI yang bertanggung jawab.
4. **Tenaga kerja:** Rekrut, kembangkan, dan pertahankan personel dengan keahlian dan pengalaman multidisiplin dalam perancangan, pengembangan, dan pemantauan sistem AI.
5. **Keterlibatan pemangku kepentingan**: Libatkan beragam perspektif dari komunitas pemangku kepentingan di seluruh siklus hidup AI untuk memitigasi risiko.
6. **Manajemen risiko**: Terapkan rencana manajemen risiko untuk mengidentifikasi, menganalisis, dan memitigasi risiko secara sistematis.

Di tingkat sistem, tata kelola mengadvokasi kepatuhan sistem AI terhadap kerangka regulasi. Berbeda dengan tingkat organisasi, yang berfokus pada praktik dan proses organisasi, tata kelola sistem menargetkan kepatuhan sistem AI dalam konteks penggunaannya, sebagaimana didefinisikan oleh peraturan dan perundang-undangan yang relevan. Dalam hal ini, kerangka kerja menetapkan tiga praktik yang direkomendasikan:

1. **Spesifikasi:** Menetapkan dan mendokumentasikan spesifikasi teknis untuk memastikan bahwa sistem AI memenuhi tujuannya.
2. **Kepatuhan**: Memastikan bahwa sistem AI mematuhi undang-undang, peraturan, standar, dan pedoman yang relevan.
3. **Transparansi:** Mendorong transparansi agar pemangku kepentingan eksternal dapat mengakses informasi tentang desain, operasi, dan batasan sistem AI.

Kerangka kerja menyediakan daftar pertanyaan untuk dipertimbangkan dan serangkaian prosedur audit untuk setiap praktik. Elemen-elemen ini krusial untuk mengoperasionalkan kerangka kerja. Gambar 6.1 dan 6.2 menunjukkan contoh pertanyaan dan prosedur audit untuk setiap praktik tata kelola.

Data: Kerangka kerja ini menetapkan bahwa data yang digunakan untuk melatih, memvalidasi, dan menguji sistem AI harus sesuai untuk memastikan sistem menghasilkan hasil yang konsisten dan akurat. Manajemen data harus menjamin kualitas, keandalan, dan representasi data secara memadai. Untuk memenuhi tujuan ini, kerangka kerja ini membedakan antara data yang digunakan untuk pengembangan model dan data yang digunakan untuk operasi sistem. Praktik yang disarankan untuk pengembangan sistem meliputi:

1. **Sumber**: Dokumentasikan asal sumber data yang digunakan untuk melatih sistem (asal data).
2. **Keandalan**: Data yang digunakan untuk melatih sistem harus andal, karena keandalan data memengaruhi akurasi hasil.
3. **Kategorisasi**: Tentukan atribut yang digunakan untuk mengkategorikan data, dokumentasikan alasan yang digunakan untuk mengorganisasikan data dan bagaimana data tersebut disegmentasi menjadi partisi pelatihan, validasi, dan pengujian.
4. **Pemilihan Variabel**: Tentukan variabel yang digunakan untuk membangun setiap komponen sistem AI. 5. Peningkatan: Tentukan penggunaan data sintetis, imputasi, atau augmentasi.

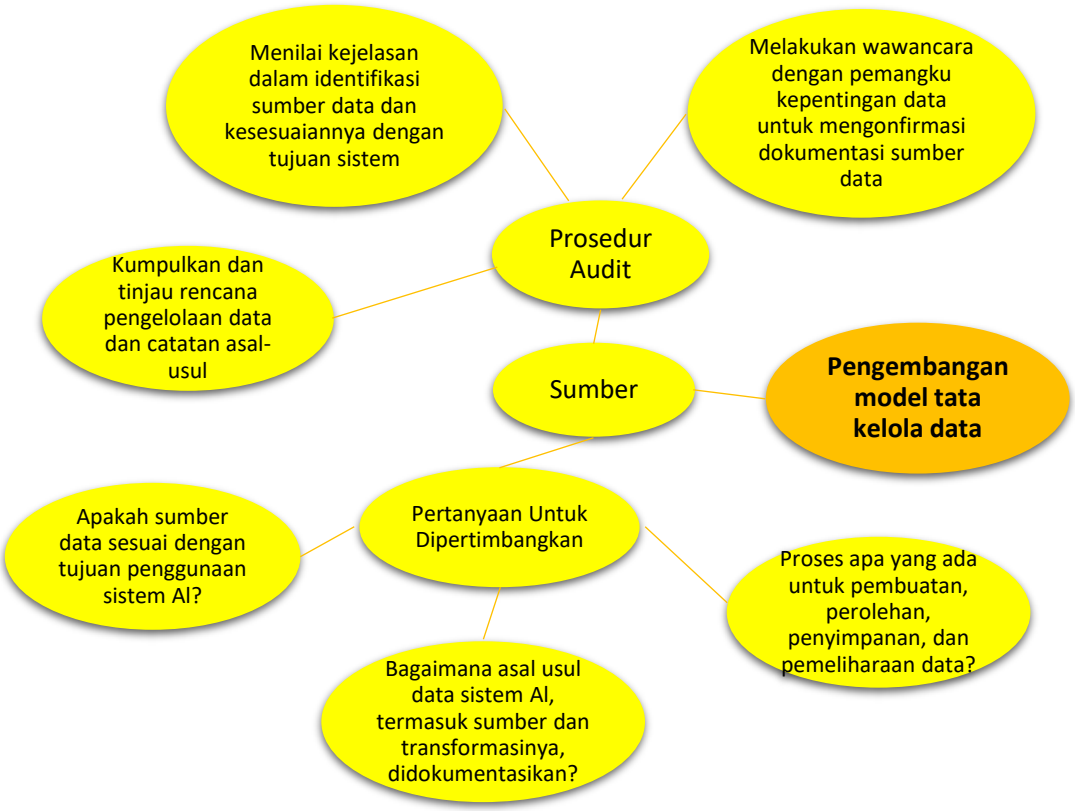
Mengenai data yang digunakan untuk operasi sistem, kerangka kerja ini menguraikan praktik-praktik berikut:

Gambar 6.1: Prosedur audit dan pertanyaan yang perlu dipertimbangkan terkait spesifikasi dalam audit AI.

Gambar 6.2: Prosedur audit dan pertanyaan yang perlu dipertimbangkan terkait kepatuhan dalam audit AI.

1. **Ketergantungan:** Menentukan interkonektivitas dan ketergantungan aliran data yang mengoperasionalkan sistem AI.
2. **Keamanan dan Privasi:** Menentukan langkah-langkah keamanan dan privasi data untuk sistem AI.

Serupa dengan tata kelola, kerangka kerja ini menguraikan pertanyaan dan prosedur audit untuk setiap praktik data yang telah ditentukan. Ringkasannya ditunjukkan pada Gambar 6.3 dan 6.4.



Gambar 6.3: Prosedur audit dan pertanyaan yang perlu dipertimbangkan terkait sumber dalam audit AI.

Kinerja. Manajemen dan pihak yang bertanggung jawab atas supervisi AI harus menggunakan metode verifikasi kinerja untuk meningkatkan akurasi hasil dan proses model, sehingga memudahkan pengambilan keputusan bagi supervisor atau tindakan korektif dan berkontribusi pada akuntabilitas publik. Dalam dimensi ini, kerangka kerja mendefinisikan sembilan praktik yang dikelompokkan pada tingkat komponen dan sistem. Pada tingkat komponen, verifikasi kinerja setiap komponen didefinisikan, karena komponen merupakan blok pembangun sistem AI. Pada tingkat sistem, verifikasi kinerja menentukan komponen mana yang beroperasi dengan baik sebagai satu kesatuan yang terintegrasi. Praktik tingkat komponen meliputi:

Gambar 6.4: Prosedur audit dan pertanyaan yang perlu dipertimbangkan terkait keandalan dalam audit AI.

1. **Dokumentasi:** Katalog model dan komponen yang bukan merupakan model, beserta spesifikasi dan parameter operasional.
2. **Metrik:** Metrik kinerja yang digunakan selama fase pengembangan sistem, yang harus akurat, konsisten, dan dapat direproduksi. Metrik harus melampaui akurasi dan mencakup bias, kesetaraan, dan pertimbangan sosial lainnya. Metrik harus mencerminkan dampak sosial yang diharapkan dari sistem AI.
3. **Verifikasi:** Verifikasi kinerja setiap komponen terkait dengan metrik yang telah ditentukan untuk memastikan sistem berfungsi sebagaimana mestinya.
4. **Keluaran:** Verifikasi keluaran mana dari setiap komponen yang sesuai dengan konteks operasional sistem AI.

Ringkasan pertanyaan dan prosedur utama terkait kinerja audit AI pada tingkat komponen ditunjukkan pada Gambar 6.5, 6.6, 6.7, dan 6.8. Praktik tingkat sistem meliputi:

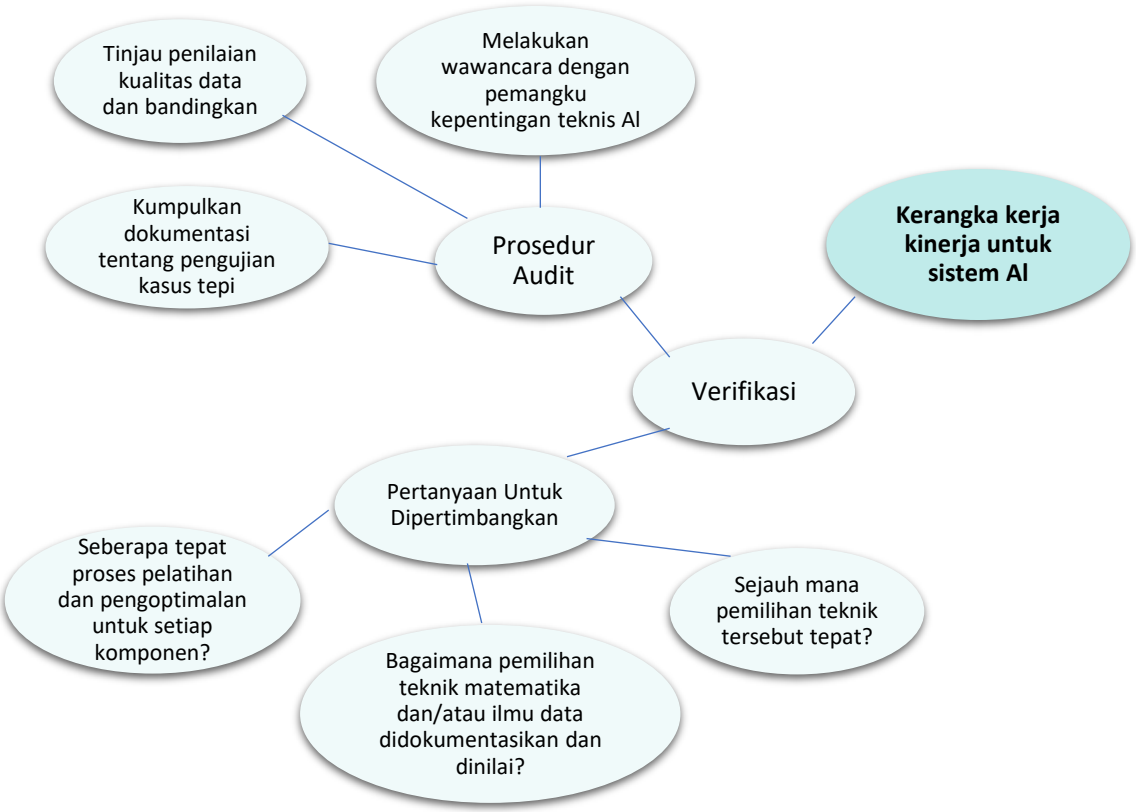
1. **Dokumentasi**: Dokumentasikan metode verifikasi, metrik kinerja, dan keluaran sistem AI untuk memberikan transparansi.
2. **Metrik**: Tentukan metrik yang akan digunakan untuk evaluasi tingkat sistem.
3. **Verifikasi**: Verifikasi kinerja terkait metrik yang ditentukan untuk memastikan sistem AI tangguh terhadap segala upaya penyalahgunaan sistem.
4. **Bias**: Identifikasi potensi bias, ketidakadilan, dan masalah sosial lainnya yang diakibatkan oleh sistem AI.

5. **Supervisi Manusia**: Menetapkan dan mengembangkan prosedur untuk supervisi manusia terhadap sistem AI yang menjamin akuntabilitas.

Kerangka kerja ini menekankan pentingnya supervisi manusia dalam sistem AI. Presentasi terperinci mengenai berbagai pendekatan supervisi manusia dapat ditemukan dalam "Model AI Governance Framework, Edisi ke-2." yang dikembangkan oleh Komisi Perlindungan Data Pribadi (Singapura). Pendekatan-pendekatan ini dikategorikan menjadi tiga:

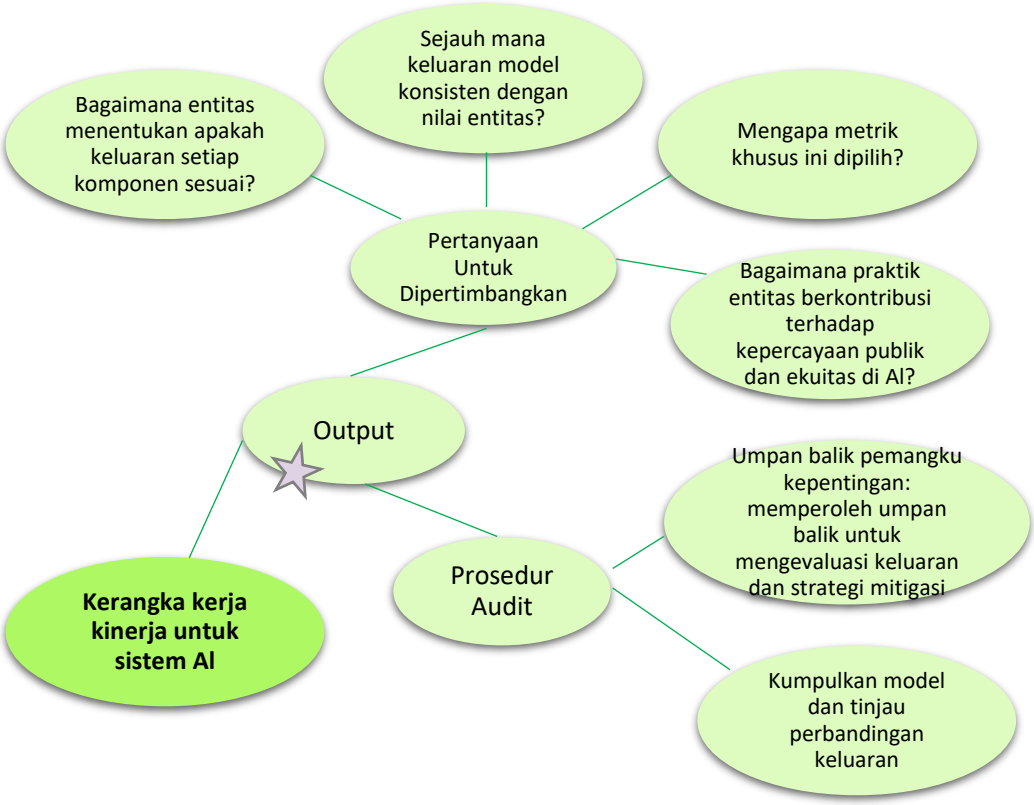
1. **Manusia-dalam-lingkaran**: Pendekatan ini melibatkan supervisi manusia aktif terhadap sistem AI, di mana manusia memegang kendali penuh atas pengambilan keputusan, dan sistem AI hanya memberikan masukan atau rekomendasi.
2. **Manusia-di-luar-lingkaran**: Ini mengacu pada tidak adanya supervisi manusia dalam pelaksanaan keputusan, di mana sistem AI memiliki kendali penuh atas pengambilan keputusan tanpa pengawasan manusia.
3. **Manusia-dalam-lingkaran**: Dalam pendekatan ini, supervisi manusia melakukan intervensi ketika hasil yang diberikan oleh sistem AI tidak sesuai dengan yang diinginkan. Jika tidak, sistem beroperasi tanpa supervisi manusia.

Gambar 6.5: Prosedur audit dan pertanyaan yang perlu dipertimbangkan terkait dokumentasi dalam audit AI di tingkat komponen.



Gambar 6.6: Prosedur audit dan pertanyaan yang perlu dipertimbangkan terkait verifikasi dalam audit AI di tingkat komponen.

Gambar 6.7: Prosedur audit dan pertanyaan yang perlu dipertimbangkan terkait metrik dalam audit AI di tingkat komponen.



Gambar 6.8: Prosedur audit dan pertanyaan yang perlu dipertimbangkan terkait keluaran dalam audit AI di tingkat komponen.

Pemantauan. Sistem AI bersifat dinamis dan adaptif, dan kinerjanya dapat berubah seiring waktu. Sangat penting untuk menetapkan kerangka kerja pemantauan guna memastikan sistem AI mempertahankan utilitasnya. Kerangka kerja ini menguraikan lima praktik untuk prinsip ini, yang dibagi menjadi kategori pemantauan berkelanjutan dan verifikasi berkelanjutan. Pemantauan berkelanjutan melibatkan pelacakan data masukan, keluaran yang dihasilkan oleh model prediktif, dan parameter kinerja yang menentukan kerangka kerja tempat sistem harus beroperasi.

Verifikasi berkelanjutan berfokus pada pemeriksaan utilitas sistem AI, terutama ketika kerangka kerja regulasi dan lingkungan operasional dapat berubah seiring waktu. Dalam beberapa kasus, entitas dapat mempertimbangkan untuk meningkatkan skala penggunaan sistem AI di berbagai lokasi geografis atau memperluas penggunaannya di berbagai pengaturan operasional. Praktik yang dikelompokkan dalam pemantauan berkelanjutan meliputi:

1. **Perencanaan**: Mengembangkan rencana untuk pemantauan berkelanjutan sistem AI guna memastikan kinerjanya sesuai harapan.
2. **Penyimpangan:** Menentukan rentang data dan penyimpangan model yang dapat diterima untuk memastikan sistem AI menghasilkan hasil yang diharapkan. Penyimpangan data mengacu pada perubahan properti statistik data masukan dalam lingkungan operasional dibandingkan dengan data yang digunakan untuk pelatihan. Penyimpangan model mengacu pada perubahan hubungan antara masukan data dan

keluaran prediksi. Penyimpangan data dan model dapat menurunkan kinerja sistem, dan rentang penyimpangan yang dapat ditoleransi harus ditentukan.

3. **Ketertelusuran:** Mendokumentasikan hasil kegiatan pemantauan dan tindakan korektif yang diambil untuk meningkatkan transparansi sistem.

Ringkasan pertanyaan dan prosedur utama untuk audit AI terkait pemantauan disajikan pada Gambar 6.9, 6.10, dan 6.11. Mengenai dimensi keberlanjutan dan perluasan penggunaan, kerangka kerja ini mendefinisikan dua praktik:

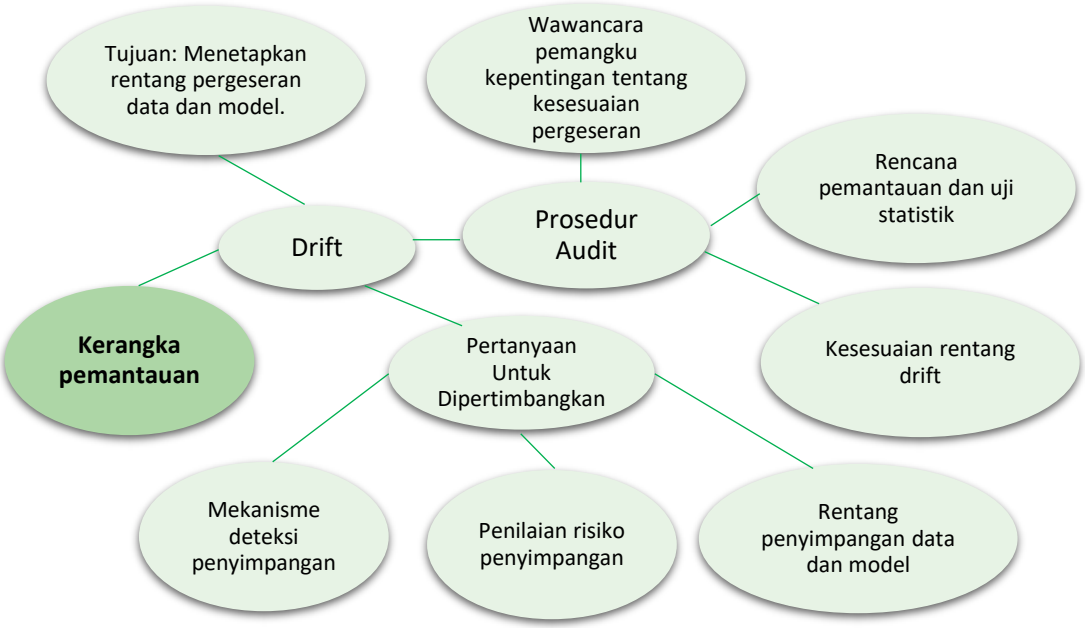
1. **Penilaian berkelanjutan**: Memverifikasi utilitas sistem untuk memastikan relevansinya dengan konteks penggunaannya. Perubahan konteks yang drastis, seperti penggunaan sistem AI selama pandemi COVID-19, memerlukan tinjauan kinerja agar sesuai dengan konteks baru.
2. **Penskalaan:** Mengidentifikasi kondisi, jika ada, yang memungkinkan sistem AI dapat ditingkatkan skalanya atau diperluas melampaui penggunaan biasanya.

Gambar 6.9: Prosedur audit dan pertanyaan yang perlu dipertimbangkan terkait pemantauan dan perencanaan dalam audit AI.

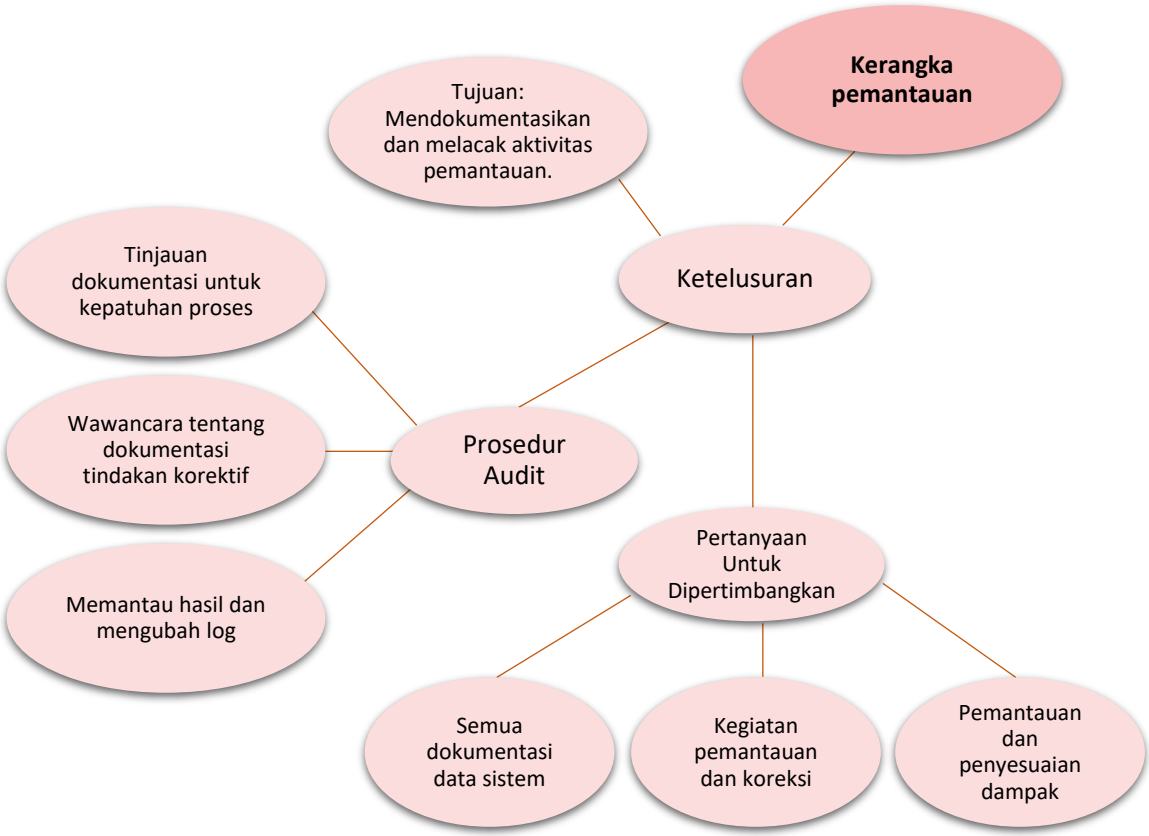
Manajemen Risiko Kecerdasan Buatan AS (US NIST AI RMF)

Pada Januari 2023, Institut Nasional Standar dan Teknologi (NIST) Departemen Perdagangan AS merilis Kerangka Kerja Manajemen Risiko AI (AI RMF) untuk membantu organisasi dalam memitigasi risiko sistem AI. Risiko ini melampaui risiko yang dibahas dalam regulasi perangkat lunak tradisional. Kerangka kerja ini menyoroti bahwa tidak seperti perangkat lunak konvensional, sistem AI dilatih berdasarkan data yang berubah seiring waktu, yang memengaruhi keandalannya dengan cara yang kompleks. Sistem AI pada dasarnya

bersifat sosioteknis, dipengaruhi oleh dinamika masyarakat dan perilaku manusia. Risiko muncul dari interaksi antara aspek teknis dan faktor sosial yang terkait dengan cara suatu sistem digunakan, interaksinya dengan sistem AI lainnya, penggunanya, dan konteks sosial tempat sistem tersebut diterapkan.



Gambar 6.10: Prosedur audit dan pertanyaan yang perlu dipertimbangkan terkait pemantauan dan penyimpangan dalam audit AI.



Gambar 6.11: Prosedur audit dan pertanyaan yang perlu dipertimbangkan terkait pemantauan dan keterlacakan dalam audit AI.

AI RMF dirancang untuk membekali organisasi dan individu dengan pendekatan yang meningkatkan keandalan sistem AI dan mendorong desain, pengembangan, penerapan, dan penggunaan yang bertanggung jawab dari waktu ke waktu. Kerangka kerja ini membahas bagaimana organisasi dapat mengatasi risiko terkait AI dan mendefinisikan karakteristik sistem AI Tepercaya, termasuk validitas, reliabilitas, keselamatan, keamanan, ketahanan, akuntabilitas, transparansi, kemudahan dijelaskan, kemudahan ditafsirkan, peningkatan privasi, dan keadilan. Inti dari kerangka kerja ini menjelaskan empat fungsi spesifik: mengatur, memetakan, mengukur, dan mengelola.

Secara spesifik, RMF AI dimulai dengan mendefinisikan bagaimana kerugian dapat memengaruhi manusia, organisasi, dan ekosistem. Kerugian terhadap manusia dapat berdampak pada individu, misalnya, dengan memengaruhi kebebasan sipil, hak, keselamatan fisik atau psikologis, atau peluang ekonomi. Kerangka kerja ini juga berfokus pada kerugian di tingkat kelompok, seperti diskriminasi terhadap kelompok tertentu, dan dampak sosial, seperti memengaruhi partisipasi dalam proses demokrasi atau akses pendidikan. Kerugian bagi organisasi dapat mencakup kerusakan pada operasi bisnis, pelanggaran keamanan, kerugian moneter, atau kerusakan reputasi organisasi.

Kerangka kerja ini juga membahas kerugian terhadap ekosistem, yang diilustrasikan oleh dampak pada elemen dan sumber daya yang saling terhubung dan saling bergantung, sistem keuangan global, sistem lain yang saling terkait, serta sumber daya alam dan lingkungan. Di sisi lain, sistem AI yang dapat dipercaya dapat mengurangi risiko dan memberikan manfaat bagi masyarakat, organisasi, dan ekosistem.

RMF AI secara efektif melengkapi kerangka kerja GAO dengan mendefinisikan tantangan dalam konteks kerangka kerja manajemen risiko. Tantangan-tantangan ini meliputi:

1. Risiko yang terkait dengan perangkat lunak, perangkat keras, dan data pihak ketiga: Risiko tidak hanya melekat pada pengembang sistem AI, tetapi juga pada penyedia layanan dan komponen yang membentuk blok pembangun sistem AI.
2. Pelacakan risiko yang muncul: Risiko baru dapat muncul karena dinamika yang menentukan perubahan dalam kondisi kontekstual tempat sistem AI beroperasi.
3. Ketersediaan metrik yang andal: Belum ada konsensus mengenai ketahanan dan verifikasi metode pengukuran risiko dan reliabilitas. Metrik yang berbeda dapat menghasilkan kesimpulan yang berbeda, sehingga hasil analisis risiko dapat saling bertentangan.
4. Risiko pada berbagai tahap siklus hidup sistem AI: Beberapa risiko mungkin laten pada tahap awal siklus hidup dan meningkat seiring perkembangan sistem.
5. Risiko dalam pengaturan dunia nyata: Pengukuran risiko dalam pengaturan laboratorium dapat menghasilkan kesimpulan yang berbeda dari yang terukur dalam pengaturan dunia nyata.
6. Ketertanam: RMF AI menggunakan konsep ketertanam untuk merujuk pada sistem tertanam, yang mana, karena sistem AI dapat menjadi bagian dari sistem yang lebih besar, hal ini memengaruhi transparansi keseluruhan sistem.

7. Dasar Manusia: RMF AI juga menekankan pentingnya mengganti manusia dengan AI, dengan memahami bahwa beberapa metrik dasar diperlukan untuk membandingkan kinerja manusia dengan kinerja sistem AI.

RMF AI mengakui kesulitan dalam mengembangkan sistem AI yang bebas risiko. Karena risiko melekat pada inovasi, RMF ini memperkenalkan konsep Toleransi Risiko, yang menunjukkan bahwa pengembangan sistem AI harus berdampingan dengan risiko dan mengelolanya untuk memitigasi dampak buruknya. Dalam hal ini, RMF AI bertujuan untuk memprioritaskan risiko dan menyadari bahwa, bahkan setelah mengelola dan memitigasi risiko, beberapa risiko residual yang melekat pada teknologi mungkin tetap ada.

Kontribusi signifikan RMF AI berfokus pada pendefinisian karakteristik sistem AI yang Tepercaya. RMF menyatakan bahwa valid dan reliabel adalah karakteristik dasar suatu sistem, yang menjadi dasar atribut-atribut lainnya. Selain itu, akuntabilitas dan transparansi merupakan karakteristik lintas sektor dari semua fitur sistem, karena keduanya berkaitan dengan bagaimana sistem AI bertanggung jawab atas suatu fitur dan atas bukti apa sistem AI mengungkapkan fitur tersebut (transparansi). Gambar 6.12 mengilustrasikan keterkaitan antara karakteristik-karakteristik ini.

Gambar 6.12: "Prinsip-prinsip AI yang Dapat Dipercaya".

RMF AI mendefinisikan masing-masing karakteristik berikut:

1. **Valid:** Ini mengacu pada konfirmasi, melalui penyediaan bukti objektif, bahwa sistem memberikan hasil yang diharapkan. Karakteristik ini berkaitan dengan akurasi sistem.
2. **Reliabel:** Kemampuan sistem untuk bekerja sesuai harapan tanpa kegagalan selama periode tertentu dan dalam kondisi operasional tertentu. Karakteristik ini berkaitan dengan ketahanan sistem.
3. **Aman**: Sistem tidak boleh menimbulkan risiko bagi jiwa manusia, kesehatan, properti, atau lingkungan.
4. **Aman dan Tangguh**: Suatu sistem dikatakan aman jika dapat beroperasi dengan tetap menjaga keandalan, integritas, dan ketersediaan melalui mekanisme perlindungan yang mencegah akses tanpa izin. Suatu sistem dikatakan tangguh jika dapat beroperasi dalam menghadapi kejadian buruk yang tidak terduga atau perubahan lingkungannya. Ketahanan berkaitan dengan kemampuan untuk menahan contoh-contoh yang merugikan, keracunan data, dan eksfiltrasi data atau informasi saat mengakses sistem.
5. **Dapat Dijelaskan dan Diinterpretasikan**: Sistem dapat dijelaskan jika menyediakan representasi mekanisme yang mendasari pengoperasian model AI. Sistem dapat diinterpretasikan jika hasil yang dihasilkan bermakna dalam konteks pengoperasian sistem.
6. **Privasi yang Ditingkatkan**: Privasi mengacu pada norma dan praktik yang melindungi otonomi, identitas, dan martabat manusia.
7. **Adil:** Sistem AI dikatakan adil jika mempertimbangkan kondisi kesetaraan dan keadilan dalam desain dan pengembangannya, serta menghindari bias dan diskriminasi yang merugikan.
8. **Akuntabel dan Transparan**: Akuntabilitas mensyaratkan transparansi. Transparansi mencerminkan sejauh mana informasi tentang sistem AI tersedia bagi pengguna. Akuntabilitas melibatkan pendefinisian prosedur dan struktur tata kelola yang menyediakan informasi ini secara tepat waktu.

Inti RMF AI

Inti RMF AI menetapkan empat fungsi untuk mengelola manajemen risiko AI. Fungsi 'Peta' mengidentifikasi konteks pengoperasian sistem AI dan risiko yang terkait dengan konteks yang diidentifikasi. Selanjutnya, fungsi 'Ukur' memastikan bahwa risiko yang teridentifikasi ditangani, dianalisis, dan dipantau. Fungsi 'Kelola' melibatkan penentuan prioritas risiko dan tindakan sesuai dengan dampak proyek. Ketiga elemen ini beroperasi berdasarkan fungsi keempat, 'Tata Kelola', yang membangun budaya manajemen risiko. Setiap fungsi AI RMF Core mendefinisikan kategori dan subkategori. Berdasarkan kategori-kategori ini, fungsi-fungsi tersebut dioperasionalkan dalam AI RMF. Kategori tingkat atas untuk setiap fungsi adalah sebagai berikut.

Tata Kelola: Fungsi ini menerapkan budaya risiko dalam organisasi, yang tercermin dalam proses, dokumen, dan skema organisasi untuk mengantisipasi, mengidentifikasi, dan mengelola risiko sistem AI. Praktik yang terkait dengan fungsi ini meliputi:

- **Kebijakan, proses, prosedur, dan praktik dalam organisasi terkait pemetaan, pengukuran, dan pengelolaan risiko AI**: Ini mencakup pemahaman persyaratan hukum dan peraturan, integrasi fitur AI yang tepercaya, penetapan parameter operasi risiko seperti toleransi risiko organisasi, dan penetapan kebijakan, prosedur, proses, dan praktik yang transparan. Fungsi ini juga mempertimbangkan penetapan peran dan tanggung jawab untuk pemantauan dan peninjauan berkala terhadap proses.
- **Struktur Akuntabilitas**: Penetapan peran dan tanggung jawab terkait pemetaan, pengukuran, dan pengelolaan risiko AI, pelatihan personel dan mitra untuk bekerja sesuai dengan kebijakan, prosedur, dan perjanjian manajemen risiko. Terakhir, pimpinan eksekutif bertanggung jawab atas keputusan yang terkait dengan manajemen risiko.
- **Keberagaman Tenaga Kerja**: Tim pengembangan dibentuk berdasarkan kriteria seperti keberagaman demografi, disiplin ilmu, pengalaman, dan latar belakang.
- **Tim organisasi berkomitmen pada suatu budaya:** Kebijakan organisasi mendorong pemikiran kritis untuk meminimalkan potensi dampak negatif sistem AI. Dokumentasi dan komunikasi dampak didorong, dan praktik organisasi mendorong pengujian, identifikasi insiden, dan berbagi informasi.
- **Proses memfasilitasi keterlibatan yang kuat dengan para pelaku AI yang relevan:** Kebijakan organisasi mendorong pengumpulan, penentuan prioritas, dan integrasi umpan balik dari para pemangku kepentingan eksternal mengenai potensi dampak sosial terkait risiko AI.
- Kebijakan dan prosedur mempertimbangkan risiko yang dihasilkan oleh ketergantungan pada perangkat lunak, data, dan penyedia rantai pasokan pihak ketiga lainnya.

Peta: Fungsi ini menetapkan konteks untuk mengelola risiko yang terkait dengan sistem AI. Praktik yang terkait dengan fungsi ini meliputi:

- **Konteks ditentukan dan disosialisasikan**: Pengetahuan yang menyeluruh tentang kerangka peraturan dan hukum, serta potensi penggunaan sistem di mana sistem AI akan diterapkan, dipertahankan. Kemampuan dan pelaku yang digunakan untuk menentukan konteks mencerminkan keragaman demografis.
- **Kategorisasi sistem AI telah ditetapkan**: Tugas dan metode spesifik yang digunakan untuk mengimplementasikan tugas-tugas yang akan ditangani oleh sistem AI telah ditetapkan. Informasi tentang batasan sistem AI dan bagaimana hasil sistem akan digunakan dan diawasi oleh manusia didokumentasikan. Integritas ilmiah sistem didokumentasikan, termasuk aspek-aspek terkait desain eksperimen, pengumpulan dan pemilihan data, serta keandalan sistem.
- Kemampuan sistem beserta tujuan penggunaan, tujuan, serta manfaat dan biaya yang diharapkan telah dibandingkan dengan tolok ukur yang sesuai.
- Risiko dan manfaat dipetakan ke semua komponen sistem AI, termasuk komponen yang disediakan oleh pihak ketiga.

- Dampak terhadap individu, kelompok, komunitas, organisasi, dan masyarakat luas dikarakterisasi: Ini mencakup dampak berdasarkan penggunaan yang diharapkan, penggunaan sistem serupa di masa lalu, laporan insiden publik, dan umpan balik dari aktor eksternal. Praktik dan personel untuk mendukung keterlibatan dengan aktor AI yang relevan dan integrasi umpan balik terus diimplementasikan dan didokumentasikan.

Pengukuran: Fungsi ini menggunakan metode kuantitatif, kualitatif, atau campuran untuk menganalisis, memverifikasi, dan memantau risiko AI serta dampak terkaitnya. Praktik terkait meliputi:

- **Identifikasi dan penerapan metode dan metrik**: Pendekatan dan metrik untuk mengukur risiko AI yang diuraikan dalam fungsi peta diidentifikasi dengan tepat. Risiko yang tidak terukur didokumentasikan. Relevansi metrik dan metode dievaluasi secara berkala. Pengukuran dilakukan berdasarkan keterlibatan kolaborator internal yang tidak terlibat dalam pengembangan, dan dengan konsultasi komunitas.
- **Evaluasi sistem AI untuk keandalan**: Pengujian didokumentasikan secara menyeluruh. Evaluasi yang melibatkan manusia mematuhi persyaratan etika dan mewakili populasi. Kinerja sistem AI diukur secara kualitatif atau kuantitatif dalam kondisi seperti operasi. Fungsionalitas setiap komponen sistem dipantau. Sistem AI terbukti valid dan andal. Sistem ini dievaluasi secara berkala untuk keamanan, termasuk metrik untuk mencerminkan keandalan, ketahanan, dan waktu respons sistem. Keamanan dan ketahanan sistem AI dievaluasi dan didokumentasikan secara berkala. Risiko yang terkait dengan transparansi dan akuntabilitas diperiksa dan didokumentasikan. Keterjelasan, risiko privasi, keadilan dan bias, serta dampak lingkungan juga dinilai.
- Mekanisme pemantauan berkala telah ditetapkan dan diterapkan.
- **Umpan balik mengenai efektivitas pengukuran dinilai**: Pendekatan untuk mengukur risiko AI dihubungkan dengan konteks penerapan sistem yang melibatkan pakar domain dan pengguna akhir. Hasil pengukuran mengenai kepercayaan AI dalam konteks operasional diinformasikan oleh pakar domain dan aktor terkait AI untuk memvalidasi kinerja sistem yang diharapkan. Kinerja terukur dan umpan balik dari pakar domain dan pengguna akhir diintegrasikan dan didokumentasikan.

Kelola: Fungsi ini mengalokasikan sumber daya untuk risiko yang dipetakan dan diukur secara berkala yang ditentukan oleh fungsi tata kelola. Fungsi ini mempertimbangkan rencana untuk merespons, mengomunikasikan, dan memulihkan dari insiden dan peristiwa. Praktik yang terkait dengan fungsi ini meliputi:

- **Prioritas risiko berdasarkan verifikasi dan hasil analisis lainnya dari fungsi pemetaan dan pengukuran:** Respons terhadap risiko AI diprioritaskan sesuai definisi dari fungsi pemetaan. Respons tersebut meliputi mitigasi, transfer, atau penerimaan. Risiko residual, yaitu risiko yang tidak dapat dimitigasi, didokumentasikan.
- **Strategi untuk memaksimalkan manfaat dan meminimalkan dampak negatif disiapkan, diimplementasikan, didokumentasikan, dan dikomunikasikan kepada para pelaku AI yang relevan:** Sumber daya untuk mengelola risiko AI dialokasikan, dan

prosedur diikuti untuk merespons dan memulihkan dari risiko yang tidak diketahui. Pengawasan disiapkan untuk menonaktifkan sistem AI yang menunjukkan kinerja atau keluaran yang tidak konsisten.

- **Manajemen risiko dan manfaat AI dari entitas pihak ketiga:** Risiko AI dan manfaatnya dipantau secara berkala, dan pengendalian risiko diterapkan dan didokumentasikan. Model pra-terlatih yang digunakan untuk pengembangan dipantau secara berkala.
- **Manajemen risiko, termasuk rencana respons, pemulihan, dan komunikasi didokumentasikan dan dipantau secara berkala:** Pemantauan pasca-penerapan diimplementasikan, termasuk mekanisme untuk menangkap dan mengevaluasi masukan dari pengguna dan pelaku AI terkait lainnya. Perbaikan berkelanjutan diintegrasikan ke dalam pembaruan sistem AI. Insiden dan kesalahan dikomunikasikan kepada para pelaku AI yang relevan, termasuk masyarakat terdampak.

Undang-Undang Kecerdasan Buatan dan Data (AIDA)

Pada bulan Juni 2022, Pemerintah Kanada memperkenalkan Undang-Undang Kecerdasan Buatan dan Data (AIDA) sebagai bagian dari RUU C-27, Undang-Undang Implementasi Piagam Digital 2022. Inisiatif ini bertujuan untuk menumbuhkan kepercayaan di antara pengguna terhadap teknologi digital yang memanfaatkan AI. AIDA menetapkan kerangka kerja yang dimaksudkan untuk memandu regulasi dan inovasi AI di masa mendatang. Pendekatan ini didasari oleh berbagai model internasional, termasuk Undang-Undang Kecerdasan Buatan Inggris, Kerangka Kerja Manajemen Risiko AI NIST, kerangka kerja Kantor Akuntabilitas Pemerintah, dan Undang-Undang AI Uni Eropa.

AIDA mengadopsi pendekatan berbasis risiko yang berfokus pada sistem AI berdampak tinggi, yang secara khusus disoroti oleh kerangka regulasi. Untuk menentukan apakah suatu sistem AI memenuhi syarat sebagai sistem berdampak tinggi, AIDA mempertimbangkan faktor-faktor kunci berikut:

1. Bukti risiko kesehatan dan keselamatan atau dampak buruk terhadap hak asasi manusia, berdasarkan tujuan penggunaan sistem dan potensi penyalahgunaan.
2. Tingkat keparahan potensi bahaya.
3. Skala penggunaan.
4. Sifat kerusakan yang disebabkan oleh sistem.
5. Sejauh mana secara wajar tidak memungkinkan untuk berhenti menggunakan sistem, karena alasan praktis atau hukum.
6. Disparitas bahaya yang terkait dengan faktor ekonomi, sosial, atau usia.
7. Sejauh mana potensi risiko diatur secara memadai berdasarkan undang-undang lain.

AIDA mengklarifikasi bahwa komponen, seperti model yang telah dilatih sebelumnya, tidak tunduk pada regulasi karena bukan merupakan sistem AI yang berfungsi penuh. Contoh sistem yang berfungsi penuh dan teregulasi meliputi:

1. **Sistem penyaringan yang memengaruhi akses ke layanan atau pekerjaan:** Sistem ini memberikan rekomendasi yang memengaruhi akses individu ke layanan dan pekerjaan, berpotensi menggunakan informasi diskriminatif dan menyebabkan kerugian bagi kelompok masyarakat tertentu.

2. **Sistem biometrik yang digunakan untuk identifikasi pribadi:** Beberapa sistem AI menggunakan data biometrik untuk membuat prediksi, seperti mengidentifikasi individu dari jarak jauh atau memprediksi perilaku individu.
3. **Sistem yang dapat memengaruhi perilaku manusia dalam skala besar:** Sistem rekomendasi yang disempurnakan dengan AI telah menunjukkan kapasitas untuk memengaruhi perilaku manusia secara luas. Dampak potensial dari sistem ini mencakup kerugian psikologis dan dampak pada kesehatan mental.

AIDA menetapkan persyaratan regulasi yang dirancang untuk membantu mengidentifikasi, memverifikasi, dan memitigasi risiko kerugian atau produksi hasil yang bias dalam sistem AI berdampak tinggi. Kewajiban untuk sistem ini dipandu oleh prinsip-prinsip berikut:

1. **Pengawasan dan Pemantauan Manusia**: Sistem AI berdampak tinggi harus dirancang dan dikembangkan untuk memungkinkan pengawasan manusia dalam mengelola operasi sistem, termasuk interpretasi yang tepat dari hasil sistem dalam konteksnya. Pemantauan didefinisikan sebagai aktivitas kritis yang melibatkan pengukuran dan verifikasi hasil sistem AI berdampak tinggi.
2. **Transparansi:** Prinsip ini mencakup penyediaan informasi yang memadai kepada publik mengenai dampak sistem AI.
3. **Keadilan dan Kesetaraan**: Sistem AI berdampak tinggi harus dikembangkan dengan kesadaran akan potensi dampak diskriminatifnya. Tindakan yang memadai harus diambil untuk memitigasi dampak diskriminatif tersebut terhadap individu dan kelompok.
4. **Keamanan:** Prinsip ini mengharuskan sistem AI berdampak tinggi secara proaktif mendefinisikan kerugian yang dapat ditimbulkan oleh penggunaannya, termasuk potensi penyalahgunaan sistem.
5. **Akuntabilitas:** Prinsip ini mewajibkan organisasi untuk menetapkan mekanisme tata kelola guna mematuhi semua kewajiban hukum yang terlibat dalam perancangan, pengembangan, dan penerapan sistem AI berdampak tinggi.
6. **Validitas dan Ketahanan**: Hasil dari sistem AI berdampak tinggi harus konsisten dengan tujuan sistem. Selain itu, sistem harus tangguh dalam berbagai kondisi.

AIDA menguraikan aktivitas dan langkah-langkah yang diatur untuk memitigasi risiko di setiap fase siklus hidup sistem AI. AIDA membahas empat fase berbeda:

1. **Desain Sistem:** Penilaian awal terhadap potensi risiko yang terkait dengan penggunaan sistem AI harus didefinisikan. Penting juga untuk mengidentifikasi potensi bias yang berasal dari pengumpulan data dan metode pemilihan serta menentukan tingkat interpretabilitas sistem yang diperlukan.
2. **Pengembangan Sistem:** Dataset dan model yang digunakan harus didokumentasikan. Proses evaluasi dan validasi yang ketat, termasuk pelatihan ulang jika diperlukan, harus dilakukan. Mekanisme pengawasan dan pemantauan manusia harus ditetapkan, dan penggunaan serta batasan sistem harus didokumentasikan.
3. **Penerapan Sistem:** Dokumentasi harus dipelihara untuk menunjukkan kepatuhan terhadap persyaratan desain dan pengembangan; menyediakan dokumentasi yang

sesuai kepada pengguna tentang dataset yang digunakan, batasannya, dan penggunaan sistem yang tepat; serta melakukan penilaian risiko terkait bagaimana sistem telah diterapkan.

4. **Pemantauan dan Manajemen Operasi:** Pencatatan dan pemantauan keluaran sistem harus dilakukan untuk memastikan pemantauan dan pengawasan sistem yang tepat, dengan intervensi yang diperlukan berdasarkan parameter operasional.

AIDA mendefinisikan dua jenis sanksi atas ketidakpatuhan terhadap peraturan, yang disebut pelanggaran peraturan: sanksi administratif berupa denda dan tuntutan atas pelanggaran peraturan. Selain itu, AIDA menetapkan mekanisme terpisah untuk pelanggaran pidana, yang mencakup daftar larangan:

- Perolehan dan penggunaan informasi pribadi secara ilegal untuk merancang, mengembangkan, menggunakan, atau menyediakan sistem AI. Pelanggaran privasi ini mencakup fase pelatihan model.
- Menyediakan sistem AI untuk digunakan, meskipun mengetahui bahwa sistem tersebut dapat menyebabkan kerugian serius bagi manusia atau properti.
- Mendistribusikan sistem AI dengan tujuan menipu publik dan menyebabkan kerugian ekonomi yang substansial bagi individu atau kelompok.

Kerangka kerja ini disusun untuk memastikan bahwa sistem AI dikembangkan dan digunakan secara bertanggung jawab, dengan perlindungan yang memadai untuk melindungi pengguna dan masyarakat umum.

Pagar Pengaman Kanada untuk AI Generatif

Munculnya AI Generatif (genAI) melalui aplikasi seperti ChatGPT dan Dall-e telah mendorong Pemerintah Kanada untuk mengembangkan Kode Praktik untuk jenis sistem ini. Berdasarkan diskusi dari G7 tentang risiko genAI dan inisiatif yang dikenal sebagai Proses AI Hiroshima, Kode Praktik ini bertujuan untuk memastikan bahwa pengembang, pelaksana, dan operator sistem genAI dapat memitigasi dampak berbahaya.

Elemen-elemen dasar Kode Praktik membahas aspek-aspek tertentu yang sebelumnya dicakup oleh AIDA, tetapi diperbarui dan disesuaikan secara khusus untuk genAI. Elemen-elemen ini meliputi:

1. **Keamanan:** Dalam konteks genAI, hal ini khususnya berkaitan dengan perlindungan terhadap risiko penyalahgunaan. Pengembang dan pelaksana harus mengidentifikasi cara-cara di mana suatu sistem dapat digunakan secara jahat, seperti meniru orang sungguhan. Mereka juga harus mengenali potensi penggunaan yang tidak pantas dan berbahaya, seperti menggunakan LLM untuk nasihat medis atau hukum, dan mengambil langkah-langkah untuk memitigasi risiko ini.
2. **Keadilan dan Keadilan:** Mengingat besarnya jumlah data yang digunakan untuk melatih model genAI, terdapat risiko signifikan untuk melestarikan bias dan stereotip data. Sangat penting untuk memastikan bahwa model dilatih dengan data yang tepat dan representatif. Pengembang harus memverifikasi dan mengkurasi kumpulan data untuk mencegah penggunaan data berkualitas rendah dan bias yang tidak representatif. Selain itu, pengembang, pelaksana, dan operator harus menerapkan

langkah-langkah untuk memeriksa dan memitigasi risiko keluaran yang bias, termasuk penyempurnaan model.

3. **Transparansi:** Sistem genAI menimbulkan tantangan dalam hal transparansi karena skalanya, yang mempersulit penjelasan hasilnya. Lebih lanjut, sistem ini bisa jadi tidak transparan terkait dataset yang digunakan dalam pelatihannya. Pengembang dan deployer harus menyediakan metode yang andal dan tersedia secara bebas untuk mendeteksi konten yang dihasilkan oleh sistem AI, seperti watermarking. Mereka juga harus memberikan penjelasan yang bermakna tentang proses yang digunakan untuk mengembangkan sistem, termasuk asal data pelatihan. Operator harus memastikan bahwa sistem yang berpotensi disalahartikan sebagai manusia diidentifikasi dengan jelas sebagai AI.
4. **Pengawasan dan pemantauan manusia:** Mengingat skala pengembangan dan beragamnya potensi penggunaan dan penyalahgunaan, perhatian khusus harus diberikan untuk memastikan pengawasan manusia dengan menetapkan mekanisme untuk mengidentifikasi dan melaporkan dampak negatif sebelum sistem genAI tersedia. Para pengembang dan operator harus menyediakan pengawasan manusia selama penerapan dan pengoperasian sistem, dengan mempertimbangkan skala penerapan dan bagaimana sistem akan tersedia bagi pengguna. Pengembang, pengembang, dan operator harus menerapkan mekanisme untuk mengidentifikasi dampak negatif dan melaporkannya setelah sistem beroperasi, misalnya, dengan memelihara repositori log insiden. Mekanisme mitigasi harus mengarah pada pembaruan rutin model, misalnya melalui penyempurnaan.
5. **Validitas dan Ketahanan:** Karena sistem genAI dapat digunakan dalam berbagai konteks, sistem ini rentan terhadap serangan dan penyalahgunaan. Fleksibilitas sistem genAI membutuhkan langkah-langkah ketat untuk mencegah konsekuensi yang tidak terduga. Pengembang harus menggunakan beragam metode pengujian di seluruh spektrum tugas dan konteks di mana sistem dapat digunakan, termasuk pengujian adversarial, untuk mengukur kinerja sistem dan mengidentifikasi kerentanan. Mereka juga harus menerapkan langkah-langkah keamanan siber yang tepat untuk mencegah atau mengidentifikasi serangan adversarial pada sistem, seperti melalui keracunan data.
6. **Akuntabilitas:** Sistem genAI harus dikembangkan dalam konteks organisasi yang mengakui pentingnya proses manajemen risiko multifaset, memastikan semua anggota organisasi memahami peran mereka dalam proses ini. Pengembang, deployer, dan operator harus memastikan adanya beberapa lini pertahanan, dengan mendorong audit internal dan eksternal terhadap sistem mereka, sebelum dan sesudah deployment. Mereka juga harus mengembangkan kebijakan, prosedur, dan pelatihan yang tepat waktu untuk memastikan bahwa peran dan tanggung jawab dalam proses manajemen risiko didefinisikan dengan jelas dan staf memahami kewajiban mereka.

Perbandingan kerangka kerja regulasi dan standar

Kerangka kerja yang dibahas menunjukkan persamaan dan perbedaan. Kerangka kerja GAO berpusat pada manajemen risiko, yang mendefinisikan empat prinsip utama: Tata Kelola, Data, Kinerja, dan Pemantauan. Untuk setiap prinsip, praktik spesifik diuraikan untuk mengoperasionalkan prinsip-prinsip tersebut di berbagai tahap siklus hidup sistem AI. Hal ini selaras dengan Standar Pengendalian Internal Federal dan menetapkan kerangka kerja umum untuk manajemen risiko. Selain itu, dokumen ini juga menetapkan pertanyaan untuk organisasi dan prosedur audit AI, dengan fokus pada praktik organisasi dan inisiatif audit AI, sehingga berfungsi sebagai kerangka kerja dengan panduan verifikasi.

RMF AI dari NIST AS mengadopsi pendekatan operasional, yang memungkinkan organisasi menerapkan praktik yang selaras dengan manajemen risiko sistem AI. RMF ini menguraikan empat fungsi yang harus dipenuhi organisasi: Tata Kelola, Pemetaan, Pengukuran, dan Pengelolaan, yang selanjutnya membagi fungsi-fungsi ini ke dalam subkategori di berbagai tingkat organisasi dan tahap siklus hidup sistem AI. Fokus di sini adalah pada proses organisasi.

Pendekatan Eropa berakar kuat pada GDPR. Dari inisiatif ini, kami mengidentifikasi dua pendekatan: satu dari Inggris (AIA ICO UK) yang menekankan GDPR dalam kaitannya dengan sistem AI, khususnya menyoroti aspek privasi dan perlindungan data pribadi. Pendekatan kedua yang didorong oleh Parlemen Uni Eropa di bawah Komisi Eropa (Undang-Undang AI Uni Eropa) menangani tantangan ini secara lebih luas. Kerangka kerja regulasi Uni Eropa beroperasi berdasarkan sistem dan hasil, mencantumkan larangan, dan kemudian mengatur apa yang disebut sistem AI berisiko tinggi, sebuah daftar yang didefinisikan secara khusus yang harus mematuhi regulasi. Tidak seperti pendekatan AS, yang berfokus pada sistem apa pun yang melibatkan AI, Uni Eropa secara khusus menargetkan sistem berisiko tinggi.

Pendekatan Kanada, AIDA, menetapkan kerangka regulasi yang menggabungkan sanksi administratif dan tindak pidana. Kerangka ini mempertimbangkan elemen-elemen dari Undang-Undang AI Uni Eropa dan AI RMF, dengan fokus pada sistem berdampak tinggi—daftar sistem yang didefinisikan namun tidak lengkap yang diatur berdasarkan potensi dampak dan kerugian bagi individu, kelompok, atau masyarakat luas. Berbeda dengan Undang-Undang AI Uni Eropa yang berbasis risiko, AIDA bersifat spesifik dampak, juga menggabungkan elemen-elemen AI RMF seperti manajemen risiko dan langkah-langkah mitigasi, serta memberikan contoh tindakan korektif yang memungkinkan. Pendekatan ini mengoperasionalkan prinsip-prinsip di tingkat organisasi.

Terdapat pula perbedaan konsep dari proposal, regulasi, dan standar ini. Konsep AI Tepercaya berlaku di Uni Eropa, sedangkan AI yang Bertanggung Jawab merupakan konsep yang dominan di AS. Bagian selanjutnya akan membahas lebih lanjut konsep ini, yang telah diadopsi secara luas di antara perusahaan-perusahaan teknologi besar yang terlibat dengan AI.

6.4 AI YANG BERTANGGUNG JAWAB

Sebuah konsep yang muncul dalam beberapa kerangka kerja yang diulas di atas adalah AI yang Bertanggung Jawab. Konsep ini lebih umum digunakan dan ditetapkan di lembaga-lembaga AS, tetapi juga disebutkan dalam AIDA. AI yang Bertanggung Jawab mencakup serangkaian praktik dan pendekatan yang mendukung pengembangan, penerapan, dan pengoperasian sistem AI sekaligus melindungi dari risiko dan potensi dampak negatifnya terhadap masyarakat.

Sebaliknya, Undang-Undang AI Uni Eropa membahas konsep AI yang Dapat Dipercaya, yang berfokus pada keandalan sistem, proses, dan pemantauan risiko berkelanjutan terhadap sistem AI. Meskipun terdapat kesamaan di antara konsep-konsep ini, AI yang Bertanggung Jawab lebih menekankan praktik manajemen risiko. Konsep AI yang Bertanggung Jawab diadopsi secara luas oleh perusahaan-perusahaan teknologi. Kami akan meninjau praktik-praktik ini dan kemudian membandingkannya untuk mendapatkan gambaran yang jelas tentang bagaimana konsep ini dipatuhi oleh perusahaan-perusahaan teknologi besar di seluruh dunia.

Google: Praktik AI yang Bertanggung Jawab

Google menguraikan enam praktik yang terkait dengan AI yang Bertanggung Jawab:

1. **Gunakan pendekatan desain yang berpusat pada manusia**: Sangat penting untuk mempertimbangkan bagaimana pengguna akan berinteraksi dengan sistem. Merancang fitur dengan pengungkapan bawaan yang tepat dapat menghasilkan kejelasan dan kontrol yang lebih baik, sehingga meningkatkan pengalaman pengguna. Untuk menangani beragam calon pengguna, sistem AI harus beroperasi berdasarkan daftar respons potensial, alih-alih hanya menyediakan satu respons. Umpan balik negatif yang potensial harus dimodelkan sejak awal proses desain, termasuk menentukan pengujian langsung. Selain itu, mengintegrasikan umpan balik pengguna sebelum dan selama pengembangan proyek sangatlah penting.
2. **Identifikasi beberapa metrik untuk menilai pelatihan dan pemantauan**: Menggunakan berbagai metrik, alih-alih hanya satu, akan membantu memahami trade-off antara berbagai jenis pengalaman. Pertimbangkan metrik yang mencakup umpan balik dari survei pengguna, metrik kuantitatif yang memantau kinerja sistem secara keseluruhan, dan metrik yang mencakup kesehatan produk jangka pendek dan jangka panjang, mulai dari rasio klik-tayang hingga nilai umur pelanggan. Lebih lanjut, metrik kinerja harus dipilah untuk berbagai kelompok pengguna.
3. **Jika memungkinkan, periksa data mentah Anda secara langsung**: Pertimbangkan aspek-aspek data dalam hal privasi, nilai yang hilang, label yang salah, representasi, kemiringan kinerja pelatihan/pengujian, fitur yang redundan, kesenjangan antara label proksi dan kategori aktual, serta bias data.
4. **Pahami keterbatasan dataset dan model Anda**: Model yang berbasis korelasi tidak boleh digunakan untuk inferensi kausal. Komunikasikan kepada pengguna keterbatasan model yang dikondisikan oleh dataset pelatihan. Ketergantungan ini

dapat memengaruhi kemampuan generalisasi model dan dengan demikian menyebabkan hasil yang tidak akurat dalam kasus penggunaan baru.

5. **Uji, uji, uji**: Belajar dari praktik pengujian rekayasa perangkat lunak akan membantu memastikan bahwa sistem AI beroperasi seperti yang diharapkan. Pertimbangkan praktik pengujian termasuk uji unit, uji integrasi, deteksi penyimpangan input, penggunaan dataset standar emas untuk menguji sistem, penerapan pengujian iteratif, dan penerapan prinsip-prinsip kualitas rekayasa seperti prinsip poka-yoke (kendala pembentuk perilaku).
6. **Pemantauan dan pembaruan berkelanjutan setelah penerapan**: Pemantauan berkelanjutan akan memastikan bahwa sistem memperhitungkan kinerja dunia nyata dan umpan balik pengguna. Contoh umpan balik pengguna untuk pemantauan berkelanjutan meliputi survei pelacakan kebahagiaan (HaTS) dan kerangka kerja HEART. HaTS, yang diperkenalkan oleh Google, adalah pengukuran sikap dan pengalaman pengguna dalam produk berskala besar berdasarkan pengumpulan data sikap. Di sisi lain, HEART mendefinisikan serangkaian metrik yang berpusat pada pengguna yang dirancang oleh Google untuk mengukur UX dalam aplikasi web.

Selain enam praktik umum untuk AI yang Bertanggung Jawab, Google menyoroti pentingnya empat faktor yang menantang:

1. **Keadilan**: Keadilan adalah isu yang kompleks. Model pembelajaran mesin didasarkan pada data yang mencerminkan dunia sebagaimana adanya, bukan sebagaimana seharusnya. Model-model ini dapat memperkuat bias yang merugikan. Membangun sistem yang adil di semua situasi dan budaya juga menantang. Tidak ada definisi keadilan yang standar, dan bahkan dalam situasi yang sederhana, orang dapat berbeda pendapat tentang apa yang dianggap adil. Google menekankan empat praktik terkait keadilan yang dapat membantu mengatasi tantangan ini:
    (a) Rancang model Anda dengan tujuan konkret untuk keadilan dan inklusi.
    (b) Gunakan set data representatif untuk melatih dan menguji model Anda.
    (c) Periksa sistem untuk bias yang tidak adil.
    (d) Analisis kinerja.
2. **Interpretabilitas:** Perumusan pedoman, praktik terbaik, dan alat yang bertanggung jawab secara konsisten meningkatkan kemampuan kita untuk memahami, mengendalikan, dan men-debug sistem AI. Google menekankan praktik-praktik berikut untuk meningkatkan interpretabilitas sistem AI:
    (a) Rencanakan pendekatan terhadap interpretabilitas di awal, selama, dan setelah merancang dan melatih model.
    (b) Jadikan interpretabilitas sebagai bagian inti dari pengalaman pengguna dengan melakukan iterasi bersama pengguna selama siklus pengembangan dan menyempurnakan asumsi kita tentang kebutuhan pengguna. Izinkan pengguna untuk melakukan analisis sensitivitas mereka sendiri jika diperlukan.

(c) **Rancang model agar dapat diinterpretasikan:** Gunakan model paling sederhana yang memenuhi tujuan kinerja Anda. Pelajari hubungan kausal, alih-alih sekadar korelasi, bila memungkinkan.

(d) Gunakan metrik yang mencerminkan tujuan akhir dan relevan dengan tugas akhir.

(e) **Pahami model yang telah dilatih**: Analisis sensitivitas model terhadap berbagai masukan untuk berbagai subset contoh.

(f) **Komunikasikan penjelasan kepada pengguna model:** Berikan penjelasan yang mudah dipahami dan sesuai bagi pengguna. Jika memungkinkan, berikan penjelasan alternatif.

(g) Analisis kinerja.

3. **Privasi:** Potensi sistem AI untuk mengungkap data tersembunyi dapat diminimalkan dengan menerapkan teknik yang dikembangkan secara khusus. Google menekankan praktik-praktik berikut terkait privasi:

   (a) Kumpulkan dan tangani data secara bertanggung jawab: Usahakan untuk melatih sistem Anda tanpa menggunakan data sensitif. Jika penggunaan data sensitif memang diperlukan, usahakan untuk meminimalkan penggunaannya. Anonimkan dan agregasi data masuk menggunakan alur kerja pembersihan data praktik terbaik seperti menghapus informasi identitas pribadi (PII), outlier, dan data sensitif lainnya yang dapat dide-anonimisasi.

   (b) Manfaatkan pemrosesan pada perangkat jika sesuai: Pertimbangkan pembelajaran terfederasi untuk meningkatkan privasi sistem Anda. Jika memungkinkan, terapkan pengacakan, agregasi aman, dan operasi lain untuk meningkatkan pembelajaran pada perangkat dalam konteks terfederasi.

   (c) Lindungi privasi model ML dengan tepat: Hindari menghafal secara tidak sengaja, bereksperimenlah dengan parameter untuk minimisasi data seperti ambang batas outlier dan agregasi, dan latih model ML menggunakan teknik yang memberikan jaminan privasi.

4. **Keselamatan dan Keamanan:** Keselamatan dan keamanan merupakan tantangan karena sulit untuk memprediksi skenario di mana sistem dapat diserang. Membangun sistem yang menyediakan batasan keamanan sekaligus fleksibilitas untuk beradaptasi dengan penggunaan atau pengguna baru juga merupakan tantangan. Google menyoroti tiga praktik untuk mengatasi aspek ini:

   (a) **Identifikasi potensi ancaman terhadap sistem**: Pertimbangkan potensi insentif untuk perilaku buruk dan identifikasi konsekuensi tak terduga yang mungkin terjadi ketika sistem mengalami kesalahan.

   (b) **Kembangkan strategi untuk melawan ancaman**: Uji kinerja sistem Anda dalam situasi yang bersifat adversarial. Dalam beberapa kasus, alat seperti CleverHans dapat digunakan. Bentuk tim merah internal untuk melakukan pengujian. Tim merah adalah kelompok yang berpura-pura menjadi musuh Anda.

(c) **Terus belajar untuk tetap menjadi yang terdepan:** Selalu ikuti perkembangan teknologi terkini, khususnya yang terkait dengan pembelajaran mesin yang bersifat adversarial. Pertimbangkan bahwa kerentanan mungkin ada di berbagai titik dalam rantai pasokan ML, bukan hanya di titik awal.

META: AI harus bermanfaat bagi semua orang

META mengikuti pendekatan yang didasarkan pada pengenalan perangkat dan sumber daya untuk AI yang bertanggung jawab. META mendefinisikan lima pilar yang mendukung komitmennya terhadap AI yang Bertanggung Jawab. Prinsip-prinsip tersebut adalah:

1. Privasi dan keamanan.
2. Keadilan dan inklusi.
3. Ketahanan dan keamanan.
4. Transparansi dan kendali.
5. Akuntabilitas dan tata kelola.

Sejalan dengan pilar-pilar ini, META mengumumkan beberapa inisiatif yang menunjukkan kepatuhannya terhadap prinsip-prinsip yang telah disebutkan sebelumnya:

1. **Set Data**: Untuk mewujudkan keadilan, META telah menciptakan beragam set data untuk melatih model AI. Salah satu yang menonjol adalah set data Casual Conversations v2, yang berguna untuk melatih chatbot dan mencakup anotasi demografis untuk memfasilitasi evaluasi sistem ini terhadap bias yang merugikan. Set data lainnya adalah HolisticBias, yang dirancang untuk menilai bias generatif dalam LLM.
2. **Variance Reduction System (VRS)**: Meta telah meluncurkan sistem baru untuk memastikan iklan ditayangkan secara adil di berbagai kelompok demografis. Inisiatif ini berfokus secara khusus pada iklan yang menawarkan peluang di bidang kredit, perumahan, atau pekerjaan. Dikembangkan bekerja sama dengan Departemen Kehakiman AS dan Departemen Perumahan dan Pembangunan Perkotaan, VRS bertujuan untuk mendistribusikan iklan terkait peluang secara adil dengan menghilangkan penargetan berdasarkan jenis kelamin, usia, atau kode pos. Lebih lanjut, Meta membandingkan audiens aktual dari iklan tertentu dengan audiens yang dipilih oleh pengiklan, menggunakannya sebagai ukuran luring untuk meminimalkan deviasi antara jumlah tayangan iklan yang sebenarnya dan audiens yang lebih luas yang memenuhi syarat untuk melihatnya.
3. **Asosiasi**: META menekankan pentingnya membentuk tim dan kemitraan interdisipliner yang mencakup organisasi hak-hak sipil, tim teknik, kelompok riset AI, serta tim kebijakan dan produk. Melalui kolaborasi ini, META menyempurnakan basis pengetahuan topik-topik yang diminati untuk digunakan dalam mitigasi lanjutan yang menargetkan asosiasi bermasalah secara lebih tepat.
4. **Umpan dan rekomendasi berbasis AI**: META telah menerapkan kontrol seperti "tampilkan lebih banyak / tampilkan lebih sedikit" untuk memberi pengguna kontrol yang lebih besar atas rekomendasi yang mereka terima. Misalnya, memilih "tampilkan lebih banyak" pada sebuah kiriman akan meningkatkan skor peringkatnya dan konten

serupa, sementara "tampilkan lebih sedikit" akan menurunkannya. Selain itu, Instagram telah memperkenalkan fitur untuk mengontrol umpan, seperti "favorit", yang menampilkan kiriman terbaru dari daftar akun tertentu, dan "mengikuti", yang hanya menampilkan umpan dari orang-orang yang diikuti pengguna.

5. **Kartu Sistem**: Inisiatif dokumentasi ini bertujuan untuk menjelaskan cara kerja suatu sistem. Meta telah merilis kartu sistem untuk berbagai sistem [4], termasuk sistem genAI multimodal, sistem gen AI untuk teks atau gambar, dan algoritma pemeringkatan. META juga berfokus pada Kartu Model, sebuah cara standar untuk mendokumentasikan dan memantau masing-masing model ML dengan tata kelola, akuntabilitas, dan transparansi yang konsisten. Meta menyoroti kasus-kasus di mana kartu model telah diterapkan pada penerjemahan mesin, deteksi objek fesyen, dan pengenalan ucapan bahasa Inggris. Terakhir, pentingnya Kartu Metode ditekankan, yang bertujuan untuk memungkinkan reproduktifitas model dan, akibatnya, pengenalan adaptasi terhadap model tersebut.
6. **Pendekatan Kebijakan**: Meta menyoroti perlunya menguji pendekatan kebijakan baru terhadap transparansi, keterjelasan, dan tata kelola AI berdasarkan Open Loop, sebuah inisiatif strategis global yang menghubungkan pembuat kebijakan dan perusahaan teknologi untuk mengembangkan rekomendasi kebijakan berbasis bukti.

Amazon dan AI yang Bertanggung Jawab

Seperti META, Amazon juga mengadopsi pendekatan yang berpusat pada penerapan alat dan sumber daya untuk AI yang bertanggung jawab [6]. Berlandaskan dimensi inti keadilan, kemudahan dijelaskan, privasi, keamanan, ketahanan, tata kelola, dan transparansi, Amazon menyoroti berbagai kasus penggunaan yang menerapkan prinsip-prinsip ini. Ini mencakup teknologi moderasi konten, aplikasi AI percakapan, verifikasi identitas, dan personalisasi, di antara yang lainnya.

Amazon menekankan urgensi untuk mengatasi tantangan AI generatif dalam konteks AI yang Bertanggung Jawab. Michael Kearns dan Aaron Roth, menunjukkan bahwa teknologi AI generatif menghasilkan konten terbuka yang bervariasi seiring percobaan berulang. Tantangan dalam menciptakan LLM yang adil adalah bahwa output bergantung pada prompt. Misalnya, jika prompt menyarankan, "Dr. Hanson mempelajari rekam medis pasien dengan saksama, lalu..." (tugas pelengkapan otomatis), wajar jika hasilnya menggunakan kata ganti pria dan wanita dengan frekuensi yang kurang lebih sama. Dr. Kearns mempertanyakan mengapa analisis yang sama tidak diterapkan pada perawat, akuntan, petugas pemadam kebakaran, atau tukang kayu, yang menyoroti masalah kelengkapan analisis. Dalam sistem terbuka, analisis akan selalu bias terhadap kategori yang ditentukan oleh mereka yang menentukan kategori mana yang akan dianalisis.

Para ahli mengidentifikasi tantangan utama berikut untuk mencapai AI Generatif yang bertanggung jawab:

1. **Toksisitas**: Kekhawatiran utama terkait AI Generatif adalah potensinya menghasilkan konten yang menyinggung, mengganggu, atau tidak pantas. Namun, mendefinisikan konten yang menyinggung cukup menantang karena seringkali berada di antara batas

tipis antara moderasi dan penyensoran konten. Apa yang dianggap toksik bergantung pada konteks dan budaya. Lebih lanjut, toksisitas seringkali tidak terwujud melalui serangan langsung, melainkan melalui mekanisme yang halus dan tidak langsung.

2. **Halusinasi**: LLM rentan terhadap halusinasi, yaitu pernyataan dan klaim yang terdengar masuk akal tetapi salah. Jenis kreativitas dalam AI Generatif ini dapat berbahaya dan bahkan tidak diinginkan.
3. **Hak Kekayaan Intelektual**: LLM terkadang menghasilkan teks yang merupakan parafrase dari data pelatihannya, sehingga menimbulkan masalah privasi dan hak kekayaan intelektual. Juga tidak jelas sejauh mana konten yang dihasilkan bersifat baru atau hanya memanfaatkan data pelatihan secara sembarangan. Kesulitan dalam membedakan antara konten baru dan penggunaan data yang dilindungi dicontohkan dalam aplikasi transfer gaya.
4. **Plagiarisme dan Kecurangan**: Kemampuan kreatif AI Generatif menimbulkan kekhawatiran dalam lingkungan pendidikan. Isu terkait plagiarisme dan kecurangan dalam konteks ini menggarisbawahi perlunya mengeksplorasi alternatif untuk melacak konten yang dihasilkan oleh AI Generatif, menggunakan teknik seperti watermarking atau strategi keterlacakan konten lainnya.
5. **Disrupsi Sifat Pekerjaan**: Efektivitas yang ditunjukkan oleh AI Generatif dalam berbagai tugas telah menimbulkan kekhawatiran tentang tergantinya profesi tertentu.

Para akademisi menyarankan beberapa solusi untuk tantangan ini. Untuk toksisitas, mereka menekankan pentingnya mengembangkan model pembatas yang mendeteksi dan menyaring konten yang tidak diinginkan dalam data pelatihan, prompt masukan, dan keluaran yang dihasilkan. Model-model ini membutuhkan data pelatihan yang dianotasi manusia, yang di dalamnya berbagai jenis dan tingkat toksisitas atau bias diidentifikasi. Untuk mengurangi halusinasi, langkah awal yang penting adalah mengedukasi pengguna tentang cara kerja AI Generatif, memastikan mereka memahami bahwa tidak semua data atau referensi yang disediakan oleh AI Generatif dapat diandalkan.

Strategi lain untuk mengurangi halusinasi melibatkan penerapan strategi RAG, yang mencakup informasi konteks yang dikurasi baik dalam prompt maupun selama proses pembuatan, yang menghubungkan LLM ke sumber data yang andal. Mengenai kekayaan intelektual, Dr. Kearns menyoroti strategi yang menggabungkan perspektif teknologi dengan mekanisme hukum. Dari sudut pandang teknologi, privasi diferensial dicatat sebagai pendekatan utama, di samping teknik sharding data, yang melibatkan partisi data pelatihan menjadi bagian-bagian kecil untuk membangun submodel yang kemudian digabungkan untuk menciptakan model global.

Dalam hal plagiarisme, ia menyoroti penggunaan teknik watermarking teks, yang didasarkan pada pembagian daftar kata yang akan disampel menjadi dua. Meskipun seorang LLM mungkin memilih untuk mengambil sampel dari salah satu daftar, manusia tidak akan dapat melakukan hal yang sama, sehingga pembatasan berbasis kosakata menjadi bentuk watermarking teks yang mudah dideteksi oleh algoritma.

Microsoft dan AI yang Bertanggung Jawab

Microsoft telah mendeklarasikan kebijakan kepatuhan terhadap AI yang aman, terjamin, dan tepercaya, sebagai bagian dari komitmen AI Sukarela Gedung Putih. Kebijakan ini dibangun berdasarkan tiga pilar utama: keselamatan, keamanan, dan kepercayaan. Aspek kunci dari perjanjian ini adalah komitmen terhadap Kerangka Kerja Manajemen Risiko AI (AI RMF) NIST. Microsoft berjanji untuk menerapkan kerangka kerja ini di seluruh perusahaannya. Rincian perjanjian yang ditandatangani dengan Gedung Putih mencakup pembangunan sistem AI yang aman berdasarkan evaluasi, verifikasi, dan validasi yang kuat, pengamanan penggunaan sistem AI Microsoft untuk model yang berkemampuan tinggi, dan peningkatan kepercayaan sistem AI Microsoft. Rincian lebih lanjut mengenai komitmen ini dapat diulas di.

Microsoft juga menyoroti inisiatif penelitian yang terkait dengan akademisi dan tim Human-AI-nya. Di antaranya adalah inisiatif RealML, serangkaian aktivitas terpandu yang dirancang untuk membantu peneliti ML mengenali, mengeksplorasi, dan mengartikulasikan keterbatasan yang ditemui dalam penelitian mereka. Alat ini mencakup panduan instruksional dan dokumen lembar kerja yang dapat diedit untuk mendokumentasikan dan menilai keterbatasan ini. Selain itu, karya berjudul “How different groups prioritizeeth values for Responsible AI” melaporkan sebuah survei yang mengkaji bagaimana individu memandang dan memprioritaskan nilai-nilai AI di tiga kelompok: (1) pekerja kerumunan, (2) praktisi AI, dan (3) sampel representatif populasi AS. Temuan ini menunjukkan bahwa praktisi AI menganggap nilai-nilai AI yang bertanggung jawab kurang penting dibandingkan warga negara AS. Lebih lanjut, responden perempuan dan kulit hitam yang mengidentifikasi diri sendiri memandang nilai-nilai AI yang bertanggung jawab lebih krusial dibandingkan kelompok lain, dan partisipan yang cenderung liberal lebih cenderung memprioritaskan keadilan.

Karya lain yang disorot adalah “Investigasi Kinerja dan Bias dalam Kerja Sama Tim Manusia-AI dalam Perekrutan”, yang melaporkan studi pengguna skala besar menggunakan kumpulan data biodata asli yang direkonstruksi, di mana manusia memprediksi kebenaran dasar pekerjaan kandidat dengan dan tanpa bantuan berbagai pengklasifikasi NLP. Studi ini menunjukkan bahwa model yang lebih mudah diinterpretasikan membantu mengurangi bias, sementara model yang kurang mudah diinterpretasikan justru mempertegasnya. Microsoft juga menekankan.

Pengujian Integrasi Manusia-AI (HINT), sebuah kerangka kerja berbasis komunitas untuk menguji pengalaman berbasis AI yang terintegrasi dengan alur kerja yang melibatkan manusia. HINT mendorong pengujian dalam konteks tugas pengguna realistis yang mensimulasikan pengalaman AI yang terus berkembang. Dengan mengintegrasikan pengujian selama fase pengembangan, risiko tak terduga dan skenario buruk yang tidak dipertimbangkan selama fase desain sistem dapat diidentifikasi.

Terakhir, Microsoft menekankan pentingnya kumpulan data yang bermanfaat untuk menangani deteksi toksisitas dan ujaran kebencian. Dalam konteks ini, ToxiGen adalah kumpulan data berskala besar yang dihasilkan mesin untuk menyempurnakan sistem deteksi bahasa toksik agar dapat menangani ujaran kebencian yang bersifat adversarial dan implisit untuk 13 kelompok minoritas demografis. Kumpulan data ini berisi 274 ribu pernyataan toksik

dan jinak tentang 13 kelompok minoritas, yang bermanfaat untuk penyempurnaan dan analisis model.

OpenAI dan AI yang Bertanggung Jawab

OpenAI adalah perusahaan teknologi yang telah memposisikan dirinya dengan layanan dan teknologi berbasis AI yang disruptif, terutama memanfaatkan model GenAI. Dampak signifikan ChatGPT dan Dall-e telah mendorong berbagai inisiatif penelitian yang berfokus pada manfaat dan potensi risiko yang terkait dengan GenAI. Penelitian OpenAI terutama berpusat pada apa yang mereka sebut sebagai 'penelitian penyelarasan', yang melibatkan pemberdayaan LLM dasar untuk memproses instruksi dan menangani tugas. Konsep penyelarasan model berkaitan dengan pengembangan LLM dari tugas penyelesaian teks menjadi pemahaman instruksi. Upaya penyelarasan ini menggabungkan umpan balik manusia melalui strategi pembelajaran penguatan, yang detailnya akan dibahas nanti dalam buku ini.

OpenAI menekankan peningkatan berkelanjutan dalam kemampuan sistem mereka untuk belajar dari umpan balik manusia. Dalam upaya ini, mereka bertujuan untuk memastikan sistem AI mereka mematuhi nilai-nilai kemanusiaan. Teknologi mereka berfokus pada tiga pilar utama:

1. Melatih sistem AI menggunakan umpan balik manusia.
2. Melatih sistem AI untuk membantu evaluasi manusia.
3. Melatih sistem AI untuk melakukan riset penyelarasan.

Menyelaraskan sistem AI dengan nilai-nilai kemanusiaan juga menghadirkan berbagai tantangan sosioteknis yang signifikan. Mengenai umpan balik manusia, strategi utama OpenAI adalah Pembelajaran Penguatan. Pembelajaran ini digunakan untuk menyelaraskan model yang dikenal sebagai InstructGPT, yang diturunkan dari model bahasa yang telah dilatih sebelumnya seperti GPT-3. OpenAI mencatat bahwa versi InstructGPT mereka jauh dari sepenuhnya selaras; mereka seringkali gagal memahami instruksi sederhana, meskipun terkadang dapat menangani instruksi yang kompleks.

Dikotomi ini, mampu menangani tugas-tugas kompleks tetapi gagal pada tugas-tugas yang lebih sederhana, memerlukan penelitian lebih lanjut untuk memahami dampak Pembelajaran Penguatan dan bagaimana kualitas umpan balik manusia dapat ditingkatkan. Mereka menyatakan bahwa pemantauan efek berbahaya lebih mudah dikelola melalui OpenAI API dari pada langsung pada model InstructGPT.

Mengenai pelatihan model untuk membantu evaluasi manusia, OpenAI menunjukkan keterbatasan mendasar Pembelajaran Penguatan dari umpan balik manusia: asumsinya manusia dapat mengevaluasi tugas-tugas yang dilakukan oleh sistem AI secara akurat. Asumsi ini tidak sepenuhnya akurat, karena faktor-faktor seperti bias budaya, agama, atau usia dapat memengaruhi cara kita mengevaluasi sistem AI. OpenAI mengembangkan model yang mungkin memberi tahu evaluator manusia apa yang ingin mereka dengar, alih-alih kebenaran. Untuk meningkatkan keselarasan, OpenAI mengembangkan teknik pemodelan imbalan rekursif (RRM), yang melibatkan pelatihan model untuk membantu evaluator dalam menilai model lain pada tugas-tugas yang sulit dievaluasi secara langsung oleh manusia. OpenAI mengilustrasikan konsep ini dengan contoh-contoh:

1. Mengevaluasi ringkasan buku bisa jadi rumit bagi manusia, terutama jika mereka tidak familiar dengan buku tersebut. Namun, model peringkasan teks dapat memberikan ringkasan bab kepada anotator, yang kemudian dapat mereka gunakan untuk mengevaluasi ringkasan keseluruhan buku.
2. OpenAI melatih model untuk menulis komentar kritis pada keluarannya sendiri; dalam tugas peringkasan berbasis kueri, bantuan dengan komentar kritis membantu menyoroti kelemahan yang dideteksi manusia dalam keluaran model.

Dalam hal melatih sistem AI untuk melakukan riset penyelarasan, upaya OpenAI berfokus pada memungkinkan sistem penyelarasan untuk membuat kemajuan riset penyelarasan lebih cepat dan lebih baik daripada yang dapat dilakukan manusia. Mereka menyarankan bahwa mengevaluasi penyelarasan jauh lebih mudah daripada memproduksinya, terutama ketika asisten evaluatif disediakan.

Dengan demikian, evaluator manusia harus semakin memfokuskan upaya mereka pada penilaian penyelarasan yang dihasilkan oleh sistem AI daripada melakukan penyelarasan itu sendiri. Mengembangkan model khusus dalam domain tertentu dengan kemampuan yang sebanding dengan manusia sangatlah penting. Sistem ini seharusnya lebih mudah diselaraskan daripada sistem tujuan umum. Akhirnya, OpenAI mengakui bahwa pendekatan berbasis penyelarasan memiliki keterbatasan, termasuk:

1. Riset penyelarasan kurang menekankan pentingnya riset ketahanan dan interpretabilitas.
2. Penggunaan bantuan AI untuk evaluasi berpotensi memperkuat inkonsistensi, bias, atau kerentanan yang mendasarinya dalam asisten AI.
3. Model yang kurang mumpuni yang digunakan untuk riset penyelarasan dapat berbahaya jika model dasar ini tidak diselaraskan secara memadai.

BAB 7
KECERDASAN BUATAN YANG DAPAT DIJELASKAN

7.1 PENDAHULUAN

Dalam beberapa tahun terakhir, perkembangan dan penerapan sistem pembelajaran mesin dan AI yang pesat telah memunculkan kebutuhan kritis akan keterjelasan. Sebagaimana konsep keadilan dan bias yang dibahas dalam bab-bab sebelumnya, keterjelasan tidaklah sederhana. Hal ini mencakup berbagai ekspektasi, interpretasi, dan kriteria yang bervariasi tergantung pada konteksnya, baik teknis, regulasi, maupun etika. Perbedaan-perbedaan ini menciptakan tantangan dalam menetapkan standar universal yang jelas tentang apa yang merupakan penjelasan "baik" dari proses algoritmik.

Seringkali, diskusi seputar keterjelasan algoritmik cenderung bersifat teknis. Fokus utamanya adalah merancang sistem dan model yang mampu memberikan keluaran yang dapat diinterpretasikan, sebuah solusi yang sering dirangkum dalam payung AI yang Dapat Dijelaskan (XAI). Namun, gagasan keterjelasan tidak boleh terbatas pada ranah efisiensi teknis. Untuk sepenuhnya memahami kompleksitas dari apa yang membuat sebuah penjelasan memadai, kita harus mengambil dari ilmu sosial dan filsafat sains.

Sebagaimana dikemukakan Tim Miller dalam makalahnya yang berpengaruh, memahami apa yang merupakan penjelasan yang baik memerlukan pertimbangan proses kognitif manusia, kesesuaian kontekstual, dan tujuan di balik penjelasan tersebut. Wawasan dari ilmu sosial ini membantu kita mengevaluasi apakah suatu penjelasan bermanfaat, bermakna, atau dapat dipercaya, tidak hanya dari sudut pandang teknis tetapi juga dari perspektif manusia.

Penelitian Miller secara kritis menyoroti bahwa penjelasan yang baik harus mencerminkan model kognitif yang digunakan orang untuk memahami peristiwa dan tindakan di dunia. Ia menekankan bahwa penjelasan harus selektif, kausal, dan kontrastif, menjawab pertanyaan "mengapa" dan "mengapa tidak" yang muncul dalam penalaran manusia. Lebih lanjut, ia menyarankan bahwa penjelasan harus dievaluasi berdasarkan utilitas sosialnya, seberapa baik penjelasan tersebut sesuai dengan proses komunikasi dan pengambilan keputusan yang melibatkan manusia, alih-alih hanya berdasarkan cara kerja internal mesin. Ini merupakan wawasan penting karena opasitas algoritmik, masalah "kotak hitam", bukan sekadar rintangan teknis yang harus diatasi. Ini adalah masalah multi-aspek yang terkait dengan ketidakjelasan proses, tata kelola data, dan sistem pengambilan keputusan seputar AI.

Untuk mengatasi tantangan opasitas algoritmik secara memadai, kita harus memperluas cakupan model AI yang dapat dijelaskan itu sendiri dan mempertimbangkan konteks yang lebih luas di mana model-model ini beroperasi. Pendekatan komprehensif terhadap keterjelasan mencakup upaya menjadikan sistem AI transparan dan memastikan bahwa proses yang mendasari sistem ini, seperti pengumpulan data, prapemrosesan, dan protokol pengambilan keputusan, juga dapat dijelaskan. Komplementaritas ini memungkinkan kita untuk bergerak menuju ketertelusuran dan transparansi yang lebih besar, dengan

menanamkan keterjelasan ke dalam kerangka tata kelola yang melingkupi AI. Dengan memahami keterjelasan sebagai kombinasi antara keterjelasan mesin dan keterjelasan proses, kita dapat menetapkan kriteria yang memadai untuk kepercayaan, keadilan, dan akuntabilitas, yang mendorong penggunaan teknologi AI yang lebih etis dan bertanggung jawab.

Dengan mempertimbangkan hal-hal ini, dalam bab ini, kami akan meninjau beberapa inisiatif paling signifikan yang bertujuan untuk menjelaskan cara kerja model kotak hitam. Upaya ini bertujuan untuk meningkatkan transparansi, khususnya apa yang kami sebut sebagai transparansi algoritmik sistem AI. Transparansi algoritmik melibatkan pengungkapan mekanisme algoritmik sistem AI untuk menghasilkan keluarannya. Teknik-teknik ini berfokus terutama pada model kotak hitam, dengan yang paling relevan adalah yang berbasis jaringan saraf tiruan. Nanti dalam buku ini, kami akan menekankan jenis model ini, dengan fokus pada model Transformer, model-model dominan dalam AI Generatif.

Masalah opasitas algoritmik

Kami membedakan antara model kotak putih dan kotak hitam berdasarkan kriteria berikut: Model kotak putih memungkinkan penelusuran proses input/output dengan mudah, mengungkap mekanisme algoritmik yang digunakan untuk menghasilkan suatu hasil. Pohon keputusan adalah contoh model kotak putih. Pohon keputusan membangun output berdasarkan evaluasi kondisi. Mekanisme algoritmik yang digunakan oleh pohon keputusan didasarkan pada aturan IF-THEN-ELSE.

Misalnya, pohon keputusan yang dirancang untuk mengklasifikasikan nasabah sistem perbankan sebagai layak kredit mungkin mempertimbangkan atribut seperti akumulasi modal nasabah. Aturan dalam pohon tersebut dapat menyatakan: IF customer.capital > Rp 1.661.500.000. THEN child.right, ELSE child.left. Berdasarkan aturan ini, pohon keputusan membagi nasabah menjadi dua segmen: nasabah dengan modal > Rp 1.661.500.000. dan nasabah dengan modal ≤ Rp 1.661.500.000. Nasabah dibagi menjadi dua kelompok, yaitu nasabah yang memenuhi syarat (anak.kanan) dan nasabah yang tidak (anak.kiri).

Aturan tersebut diterapkan pada contoh-contoh dengan tujuan mencapai kemurnian tinggi terkait variabel target, yang dalam hal ini adalah kelayakan kredit nasabah. Berbagai algoritma pembelajaran mesin melatih model-model ini, dengan tujuan memaksimalkan kemurnian simpul daun pohon keputusan berdasarkan partisi pelatihan. Setelah melatih model, selama fase inferensi, mekanisme penjelasan untuk nasabah tertentu terdiri dari penggabungan semua aturan yang diterapkan untuk mengklasifikasikan nasabah tersebut.

Kami menyebut jenis penjelasan ini sebagai penjelasan lokal, karena spesifik untuk suatu kasus. Dalam contoh ini, mekanisme penjelasan secara langsung menggunakan mekanisme algoritmik model untuk menjelaskan hasil bagi nasabah tertentu. Karena model secara transparan mengungkapkan variabel dan kondisi yang digunakan untuk menghasilkan hasil, kami menyebutnya model kotak putih.

Karena kompleksitasnya, model kotak hitam tidak secara langsung menyediakan mekanisme algoritmik transparan yang membantu menjelaskan mengapa model tersebut menghasilkan suatu hasil. Hal ini terjadi pada Jaringan Syaraf Tiruan, khususnya jaringan saraf dalam, yang memiliki banyak lapisan, sehingga sangat sulit untuk melacak aturan algoritmik

yang digunakan untuk menghasilkan suatu hasil. Kami menyebut karakteristik model ini sebagai opasitas algoritmik.

Opasitas algoritmik didefinisikan sebagai properti model yang menyulitkan untuk mengungkapkan mekanisme algoritmik yang digunakan untuk menghasilkan suatu hasil. Kami secara khusus berfokus pada opasitas algoritmik JST. JST, dan khususnya jaringan saraf dalam, secara algoritmik bersifat opak karena dua faktor: a) mekanisme pembangkitan keluaran dari masukan dan b) mekanisme replikasi di beberapa lapisan. Mengenai mekanisme pembangkitan keluaran dari masukan, kami katakan mekanisme ini opak karena menyulitkan untuk melacak fitur mana dari contoh yang relevan dalam menghasilkan keluaran. Mekanisme algoritmik fundamental jaringan saraf tiruan melibatkan perkalian matriks-vektor, di mana matriks menyimpan parameter lapisan, dan vektor merepresentasikan fitur dari contoh yang sedang kita proses. Misalkan $x(0)$ adalah vektor fitur yang merepresentasikan contoh dan W (1) adalah matriks parameter lapisan pertama. Operasi dasar yang dilakukan oleh jaringan saraf tiruan adalah mengalikan matriks parameter dengan vektor fitur, yang dilambangkan dengan perkalian matriks-vektor $W\ (1)x(0)^T$, di mana T merepresentasikan operasi transpos.

Matriks memproyeksikan vektor fitur ke vektor baru, yang akan kita sebut (1). Setiap komponen vektor ini dihasilkan dari perkalian titik antara setiap baris matriks parameter dan vektor fitur. Misalnya, untuk entri ke-i dari vektor s(1), yang dilambangkan sebagai s(1)[i], operasi yang menghitung nilai s(1)$[i]$ adalah perkalian vektor-vektor W (1)$[i]x(0)^T$, di mana W (1)$[i]$ merepresentasikan baris ke-i dari matriks parameter. Kemudian, nilai skalar s(1)$[i]$ bersesuaian dengan agregasi semua komponen vektor fitur. Fungsi agregasi adalah jumlah terbobot menurut parameter baris ke-i matriks parameter. Pada titik inilah kita mulai kehilangan ketertelusuran fitur $x(0)$ karena efek setiap komponen fitur diagregasi ke dalam setiap komponen vektor s(1). Lapisan pertama jaringan saraf melibatkan satu operasi lagi, yang disebut aktivasi. Setiap entri vektor s(1) melewati fungsi aktivasi. Biasanya, fungsi aktivasi adalah fungsi nonlinier, seperti fungsi logistik atau tangen hiperbolik.

Tujuan fungsi ini adalah untuk menghasilkan nilai dalam domain yang diketahui, misalnya, untuk fungsi logistik atau untuk tangen hiperbolik, dengan mempertahankan arah pertumbuhan sinyal masukan. Semakin tinggi nilai masukan, semakin tinggi nilai keluaran. Kami menyatakan fungsi aktivasi dengan O(). Ketika setiap komponen s(1) melewati O(), kita memperoleh vektor keluaran dari lapisan tersebut, yang kita sebut $x(1)$.

Dengan demikian, komponen ke-i dari $x(1)$ diberikan oleh O(s(1)$[i]$). Pada titik ini, opasitas mekanisme algoritmik jaringan saraf tiruan bahkan lebih besar, karena untuk melacak efek fitur tertentu dari vektor $x(0)$ ke $x(1)$, kita telah melewati fungsi agregasi dan fungsi aktivasi, yang biasanya nonlinier. Memahami efek komponen $x(0)$ pada $x(1)$ sangat menantang. Ini adalah mekanisme komputasi dasar dari jaringan saraf tiruan, yang kita sebut komputasi umpan-maju.

JST, khususnya jaringan saraf tiruan dalam, memanfaatkan blok penyusun fundamental ini dengan menghubungkan beberapa lapisan secara berurutan. Pelapisan ini meningkatkan opasitas model. Seperti yang akan kita bahas nanti di buku ini, model-model dominan di bidang AI Generatif melakukan operasi tambahan pada masukan melalui mekanisme self-

attention. Mekanisme ini sangat bergantung pada komputasi umpan maju yang dijalankan secara paralel, yang selanjutnya meningkatkan opasitas model.

XAI: Mencerahkan Model Kotak Hitam

Metode XAI (Explainable Artificial Intelligence) bertujuan untuk menjelaskan model kotak hitam, memberikan penjelasan atas operasinya. Kami mengadopsi taksonomi yang diusulkan oleh Speith untuk mengkategorikan berbagai pendekatan XAI. Pada dasarnya, terdapat empat konsep kunci yang perlu dipertimbangkan:

1. **Tahap:** Bergantung pada tahap di mana metode ini melakukan intervensi, strategi XAI dapat diklasifikasikan sebagai post-hoc atau ante-hoc. Pendekatan ante-hoc melibatkan pemilihan model kotak putih, yang secara inheren memberikan penjelasan berdasarkan desainnya. Contoh model tersebut adalah pohon keputusan, yang menawarkan penjelasan lokal berdasarkan konjungsi kondisi yang dipenuhi oleh contoh spesifik. Sebaliknya, pendekatan post-hoc beroperasi setelah suatu model, biasanya model kotak hitam, telah membuat inferensi. Penting untuk menilai apakah metode XAI dapat diterapkan, karena beberapa metode spesifik untuk jenis model tertentu sementara yang lain tidak bergantung pada model.
2. **Cakupan**: Bergantung pada cakupan penjelasan yang dihasilkan oleh metode XAI, metode dapat diklasifikasikan menjadi cakupan lokal atau global. Metode XAI lokal menghasilkan penjelasan yang spesifik untuk setiap contoh. Misalnya, LIME [251], yang akan kita bahas nanti, adalah salah satu metode tersebut. Metode XAI global menghasilkan penjelasan komprehensif untuk keseluruhan model, yang menunjukkan fitur model mana yang menjelaskan setiap kelas variabel target. Metode penjelasan global mencakup pentingnya fitur, antara lain.
3. **Hasil:** Berdasarkan jenis keluaran yang dihasilkan oleh metode XAI, keluaran ini dibagi menjadi model pengganti, relevansi fitur, atau contoh. Metode XAI berbasis model pengganti menghasilkan model yang berfungsi sebagai penjelasan. LIME adalah contoh dari metode tersebut, yang menyediakan model pengganti yang secara lokal mengaproksimasi model asli di sekitar ruang representasi contoh untuk menghasilkan penjelasan. Metode XAI relevansi fitur berfokus pada identifikasi fitur yang relevan untuk menghasilkan penjelasan. Terakhir, metode XAI berbasis contoh membangun penjelasan menggunakan contoh hasil yang disediakan oleh model.
4. **Fungsi:** Bergantung pada bagaimana metode XAI beroperasi, metode ini dapat didasarkan pada perturbasi atau pemanfaatan struktur. Metode XAI berbasis perturbasi menghasilkan variasi dari suatu contoh untuk mengevaluasi pengaruh variasi ini terhadap hasil. Metode pemanfaatan struktur menggunakan struktur data atau model untuk membangun penjelasan. Cara populer untuk melakukan ini adalah dengan memeriksa gradien dalam jaringan saraf tiruan, karena gradien dapat memberikan wawasan tentang pentingnya nilai masukan individual. Pendekatan lain adalah menyederhanakan arsitektur, seperti memodifikasi fungsi di dalam model. Contoh metode XAI ini, yang dikenal sebagai modifikasi arsitektur, mencakup modifikasi yang diterapkan pada jaringan konvolusional, seperti mengganti

pengumpulan maksimum dengan pengumpulan rata-rata. Kategori lain dalam fungsi melibatkan konstruksi meta-penjelasan, yang dihasilkan dengan menggabungkan penjelasan dari metode XAI lainnya.

Kategori taksonomi ini tidak saling eksklusif. Misalnya, LIME adalah metode post-hoc, model-agnostik yang menghasilkan penjelasan lokal berdasarkan model pengganti yang beroperasi melalui perturbasi. Selain itu, metode XAI dapat menghasilkan penjelasan dalam berbagai format keluaran, termasuk aturan, penjelasan berbasis teks, penjelasan visual, penjelasan campuran (visual + teks), penjelasan berbasis argumen, atau bahkan penjelasan berbasis model.

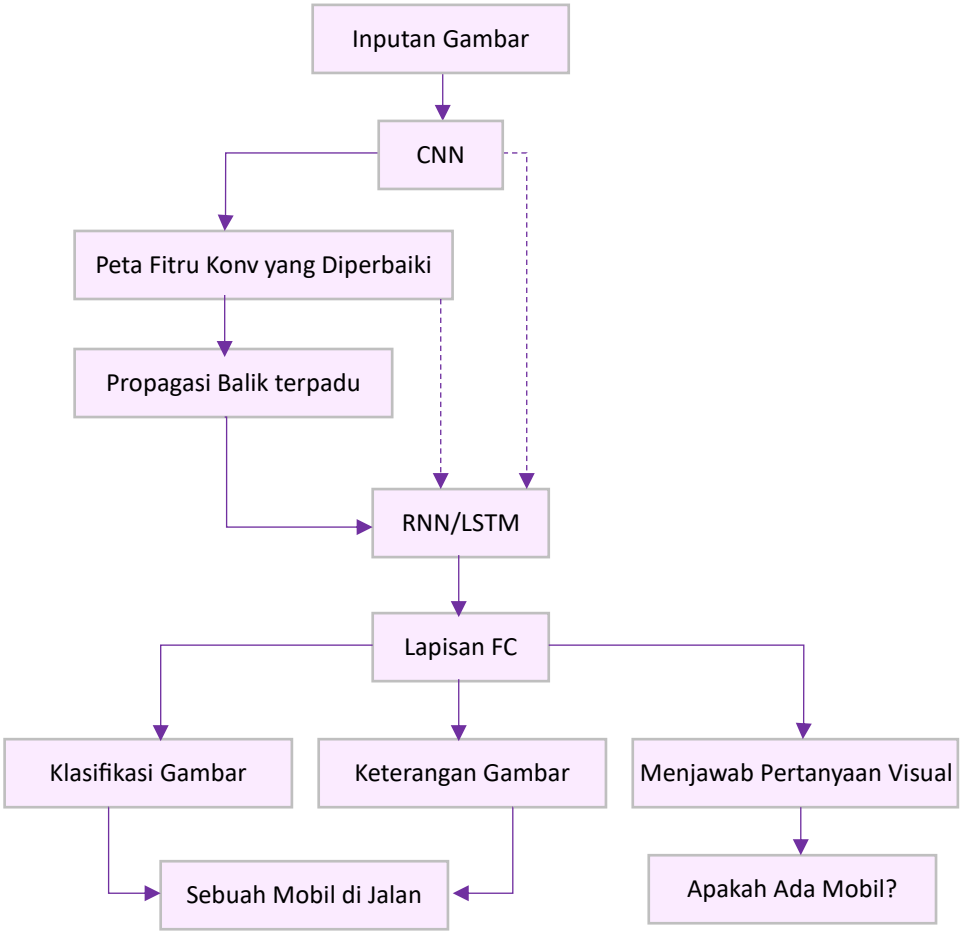
7.2 METODE XAI

XAI merupakan bidang penelitian yang aktif dengan beragam metode. Dalam konteks ini, kami akan menjelaskan tiga metode yang banyak digunakan yang menggambarkan strategi yang diterapkan untuk meningkatkan transparansi algoritmik sistem AI: LIME, Grad-CAM [269], dan Nilai Shapley.

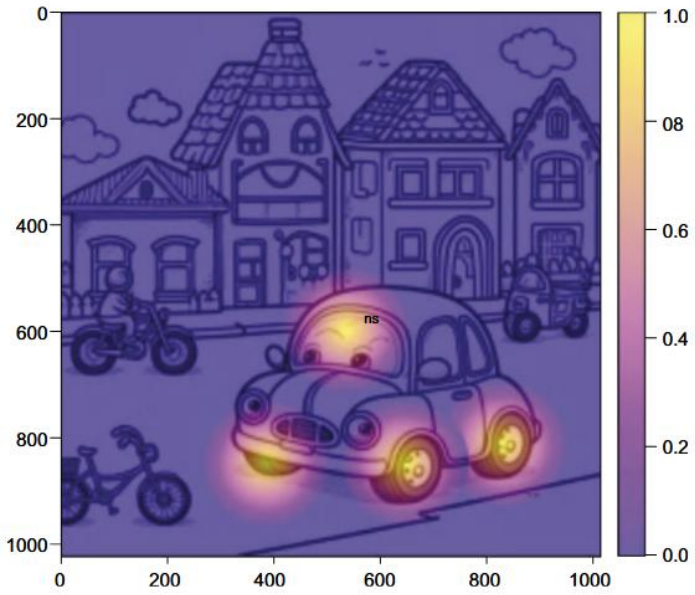
Grad-CAM

Seperti yang telah diilustrasikan di bagian sebelumnya, jaringan saraf tiruan memperkenalkan opasitas algoritmik karena desainnya menyulitkan pelacakan pengaruh setiap fitur dalam vektor masukan terhadap keluaran. Kami telah menyebutkan bahwa mekanisme pembangkitan keluaran dari masukan, yang melakukan operasi agregasi pada komponen vektor masukan dan menggunakan fungsi aktivasi nonlinier, mempersulit pelacakan pengaruh fitur terhadap keluaran. Lebih lanjut, dalam kasus jaringan saraf tiruan dalam, mekanisme agregasi ini diterapkan beberapa kali, sehingga semakin mempersulit tugas menghasilkan penjelasan.

Grad-CAM memperkenalkan teknik untuk menghasilkan penjelasan visual untuk jaringan dalam. Mekanisme ini didasarkan pada penggunaan gradien dari setiap konsep target yang mengalir ke lapisan konvolusional terakhir jaringan untuk membuat peta lokalisasi wilayah masukan yang memprediksi konsep tersebut. Grad-CAM dirancang khusus untuk jaringan konvolusional, sejenis jaringan saraf dalam yang dirancang khusus untuk memproses gambar. Ide di balik peta visualisasi Grad-CAM adalah untuk menyoroti wilayah-wilayah penting dalam gambar masukan yang mengarah ke keluaran tertentu.



Gambar 7.1: Alur data "Grad-CAM".



Gambar 7.2: Penjelasan "Grad-CAM".

Grad-CAM mengandalkan perhitungan gradien skor kelas, sebelum lapisan klasifikasi, yang biasanya diimplementasikan sebagai lapisan softmax. Gradien, yang merupakan versi multivariat dari turunan yang memungkinkan perhitungan pada vektor, dihitung sehubungan

dengan aktivasi peta fitur dari lapisan konvolusional. Gradien ini dipropagasi balik dan dirata-ratakan secara global pada dimensi lebar dan tinggi peta fitur untuk mendapatkan bobot kepentingan neuron.

Mekanisme Grad-CAM menghasilkan penjelasan lokal. Misalnya, pada suatu gambar yang objeknya terdeteksi oleh jaringan konvolusional, kita dapat melihat wilayah mana yang signifikan dalam mendeteksi suatu objek. Mekanismenya dijelaskan pada Gambar 7.1.

Misalkan, jaringan mendeteksi bahwa citra kita berisi sebuah mobil. Dengan menggunakan Grad-CAM, kita menghitung gradien skor untuk kelas 'mobil' dalam jaringan, yang terletak di lapisan konvolusional terakhir. Selanjutnya, kita menghitung peta aktivasi untuk kelas ini berdasarkan gradien tersebut. Peta-peta ini, ketika dipropagasi balik ke lapisan pertama, akan menyorot wilayah-wilayah penting pada citra. Penjelasan visual, dalam contoh ini, akan terdiri dari pengukuran korespondensi antara kelas mobil dan wilayah citra masukan. Gambar 7.2 menunjukkan contoh penjelasan visual berdasarkan Grad-CAM.

Penjelasan visual yang dihasilkan oleh Grad-CAM merupakan jenis penjelasan korespondensi. Menurut taksonomi Speith, Grad-CAM adalah metode post-hoc spesifik model yang menghasilkan penjelasan lokal dan termasuk dalam kategori struktur leverage. Metode ini spesifik model karena dirancang khusus untuk jaringan konvolusional. Metode ini post-hoc karena beroperasi selama fase inferensi. Ini bersifat lokal karena penjelasannya dihasilkan untuk satu gambar. Selain itu, ini diklasifikasikan dalam struktur leverage karena memanfaatkan fitur model untuk menyoroti fitur yang dapat dijelaskan, dalam hal ini, area yang relevan dari gambar input.

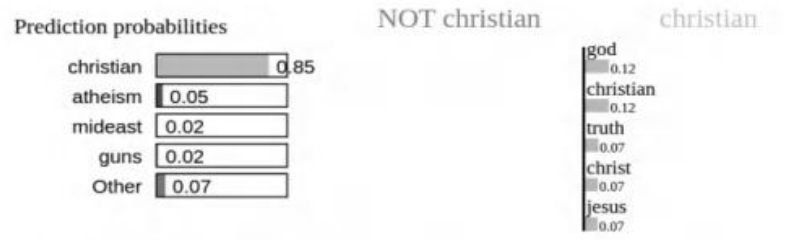
Grad-CAM menyediakan penjelasan korespondensi visual, artinya metode ini menetapkan korespondensi antara wilayah yang relevan dan kelas yang dianalisis. Metode ini bergantung pada pengguna akhir untuk merasionalisasi penjelasan tersebut dengan mengevaluasi validitas korespondensi yang terdeteksi. Misalnya, kita melihat bahwa wilayah yang disorot pada gambar berisi sebuah mobil, yang sesuai dengan kelasnya. Penjelasan ini akan masuk akal dan dengan demikian membantu kita mengevaluasi mengapa metode ini menghasilkan hasil ini. Sedikit modifikasi Grad-CAM memungkinkan untuk menghasilkan penjelasan kontrafaktual, yaitu menilai hasil dengan menghilangkan suatu wilayah dari gambar asli. Grad-CAM mencapai hal ini dengan menggunakan gradien negatif dari skor kelas, kemudian peta fitur jaringan mengidentifikasi wilayah-wilayah dari gambar masukan yang, jika dihilangkan, akan membuat model lebih yakin dengan prediksinya.

Penjelasan Model-Agnostik yang Dapat Diinterpretasikan Secara Lokal (LIME)

LIME merupakan metode terkemuka di bidang XAI. Tujuannya adalah untuk mengidentifikasi model yang dapat diinterpretasi berdasarkan representasi yang dapat diinterpretasi yang setia secara lokal, sehingga mendekati perilaku pengklasifikasi asli di dekat contoh spesifik. Untuk pengklasifikasi teks, representasi data yang dapat diinterpretasi mungkin terdiri dari vektor biner yang menunjukkan ada atau tidaknya suatu kata, meskipun pengklasifikasi tersebut mungkin menggunakan fitur yang lebih kompleks, seperti yang dikodekan dalam jaringan saraf tiruan dalam.

LIME beroperasi sebagai metode penjelasan lokal. Metode ini dimulai dengan memilih sebuah contoh yang akan menghasilkan penjelasan. Vektor fitur dari contoh ini terganggu; dalam konteks teks, gangguan ini melibatkan penghapusan sebuah kata dari dokumen. Misalnya, pertimbangkan pengklasifikasi teks yang dilatih pada dataset 20 Newsgroups, yang akan kita sebut sebagai model M. Misalkan sebuah artikel berita d diklasifikasikan oleh M dalam kategori 'Kristen'.

Tujuan kita adalah untuk memahami fitur dokumen mana, dalam hal ini, kata-kata yang memengaruhi klasifikasi ini. Untuk mencapai hal ini, LIME mengambil dokumen d dan menghasilkan versi yang terganggu dengan menghapus kata-kata. Misalkan d mencakup kata 'tuhan'. LIME akan membuat versi d yang terganggu, dengan menghapus 'god' untuk mendapatkan dokumen baru, misalnya d', dan mengklasifikasikannya dengan M. Jika $M(d')$ = 'baseball', kita menggunakan hasil ini untuk menunjukkan bahwa d': BUKAN Kristen.



Teks dengan kata yang disorot

mungkin pemerkosaan tidak ada hubungannya bekerja mungkin yang lain memberikan wawasan keseluruhan tentang **Tuhan** kebenaran secara keseluruhan agama tampaknya bergantung sepenuhnya pada individu dengan baik secara individual diciptakan karena **kristen** gagal menunjukkan cara hidup yang lebih baik sebelum percobaan upaya saya yang diperlukan bahkan sangat menarik kesejahteraan tidak pernah tahu pasti diberi tahu kebenaran setidaknya sedikit bukti yang menunjuk pada fakta mengatakan konflik militer perang vietnam setahun yang lalu dewa supernatural ingin hidup tertentu memperingatkan fakta yang benar yesus memperingatkan berarti tidak ada peringatan hal besar sebenarnya pilihan pertama dua kehidupan kematian hampir tidak terdokumentasi yang terakhir satu total omong kosong kecuali satu kita alkitab sama sekali tidak bulat mungkin penggunaan imajinasi seseorang ketidaktahuan seseorang orang lain mengatasi saya yakin lebih suka banyak dokumentasi tidak ada yang menjijikkan **kristen** percobaan memanipulasi menafsirkan perjanjian lama dipenuhi tanda kedatangan **kristus** setiap sedikit referensi tongkat sedikit akan secara otomatis ditafsirkan salib mendamaikan filsafat banyak skeptis hati berpikir syukurlah jalan perbedaan yang tidak dapat didamaikan **kristen** ketidakmampuan tepat persis cenderung menghina dosa mempercayai hanya percaya iman tanpa pengetahuan beruntung seseorang sehari terjadi berpikir **Tuhan** waktu enkefalin mungkin mengaitkan tanda **Tuhan** merasa benar mempercayai kebenaran tanpa mengetahui mungkin religiositas terlihat apa pun secara meyakinkan tiba jarang tampak penderitaan dosa mengafirmasi percaya banyak bersedia mati percaya banyak pertanyaan seperti sikap mencerminkan moralitas benar buruk tampaknya tipis dapat mencerminkan fanatisme contoh kasus ekspresi kesederhanaan keegoisan adam

Gambar 7.3: Penjelasan "LIME" untuk klasifikasi teks.

Jika kita mengganggu d lagi, kali ini dengan menghapus 'guns' untuk mendapatkan d'', dan M(d'') = 'guns', hasilnya menunjukkan d'': Kristen. Melalui mekanisme gangguan ini, kita dapat membangun himpunan data versi d yang terganggu dengan anotasi biner yang menunjukkan apakah dokumen tersebut termasuk dalam kategori 'Kristen' atau tidak. LIME akan menggunakan himpunan data ini untuk melatih pengklasifikasi biner B. Pengklasifikasi linier biasanya digunakan untuk tujuan ini karena menghasilkan pembagian ruang representasi menjadi dua segmen, yang dapat digunakan untuk membangun penjelasan. Idenya adalah dengan mengaproksimasikan M mendekati d dengan model linier B, kita dapat menggunakan B untuk menjelaskan M(d). Penjelasan yang dihasilkan menggunakan mekanisme ini diilustrasikan pada Gambar 7.3.

Grafik LIME untuk teks di atas menunjukkan tingkat keyakinan model dalam mengklasifikasikan cuplikan teks sebagai "BUKAN Kristen" atau "Kristen". Model memprediksi dengan probabilitas 85% bahwa teks tersebut "BUKAN Kristen". Dua kolom yang diberi label di bawah kategori "BUKAN Kristen" dan "Kristen" mencantumkan kata kunci yang diekstrak dari teks beserta bobotnya, yang diwakili oleh batang. Bobot ini menunjukkan seberapa besar kontribusi setiap kata terhadap prediksi model. Misalnya, kata "god" berkontribusi 0,12 terhadap prediksi "Kristen", yang berarti sedikit mendorong model untuk mengklasifikasikan teks sebagai "Kristen". Di bawah probabilitas prediksi, teks ditampilkan dengan kata-kata tertentu yang disorot. Sorotan ini sesuai dengan kata-kata yang disebutkan dalam daftar kata di atas. Visualisasi ini membantu pengguna melihat kata-kata spesifik mana dalam teks yang memengaruhi prediksi model.

Jenis penjelasan yang diperoleh dari B mengevaluasi efek setiap fitur. Mekanisme untuk membangun penjelasan ini melibatkan pengklasifikasian gangguan d menggunakan B, dan pengukuran jarak ke B. Asumsikan $B(d')$ = 0 (artinya d' bukan milik Olahraga). Jarak dari d' ke B akan menunjukkan seberapa besar pengaruh sebuah kata (kata yang dihilangkan dari d untuk menghasilkan d') terhadap hasil. Semakin besar jaraknya, semakin besar pula dampak kata tersebut terhadap klasifikasi. Jenis penjelasan yang dihasilkan oleh LIME adalah tingkat fitur, yang menunjukkan relevansi setiap fitur dalam klasifikasi. Karena B bersifat biner, salah satu cara untuk memberikan penjelasan visual adalah dengan menggunakan plot batang berdampingan. Selain itu, kata-kata dalam dokumen dapat diwarnai sesuai relevansinya dengan klasifikasi.

LIME juga dapat menghasilkan penjelasan untuk kumpulan data tabular. Prosesnya serupa dengan yang digunakan untuk teks: melibatkan perturbasi pada sebuah contoh untuk membuat kumpulan data variasi dari contoh asli, lalu menyesuaikan model linear di sekitar contoh tersebut. Hal ini dilakukan dengan menggunakan contoh-contoh yang mengalami perturbasi dan klasifikasinya sesuai dengan model asli untuk melatih pengklasifikasi baru. Perturbasi menunjukkan bobot setiap fitur dalam hasil, yang menunjukkan apakah keseimbangan klasifikasi condong ke satu sisi atau sisi lain dari bidang-hiper pemisah.

Dampak dari setiap contoh yang mengalami perturbasi, berdasarkan jaraknya ke bidang-hiper pemisah, menunjukkan relevansi fitur dalam menggoyahkan keseimbangan. Gambar 7.4 menyajikan contoh penjelasan LIME tabular pada set data diabetes, yang

mengilustrasikan fitur-fitur pasien mana yang relevan untuk mengklasifikasikan mereka sebagai sehat. Tabel tersebut juga menampilkan fitur-fitur yang berpotensi memengaruhi kondisi ke arah lain.

Gambar 7.4: Penjelasan LIME tabular pada set data diabetes mengilustrasikan fitur pasien mana yang relevan untuk mengklasifikasikan mereka sebagai sehat.

Menurut taksonomi Speith, LIME adalah metode XAI post-hoc yang dicirikan oleh beberapa fitur utama. Sebagai metode yang tidak bergantung pada model, LIME dapat mengaproksimasi model apa pun menggunakan model linier, yang berarti mekanisme aproksimasinya tidak bergantung pada model M. Secara spesifik, LIME membangun model pengganti (dilambangkan sebagai model B) untuk memfasilitasi penjelasan, menjadikannya metode XAI yang memberikan penjelasan berbasis hasil.

Dalam hal cakupan, LIME dianggap sebagai metode lokal karena model pengganti dikondisikan pada dokumen tertentu. Secara fungsional, LIME didasarkan pada perturbasi; metode ini menggunakan perturbasi untuk menghasilkan versi instans data d untuk menurunkan model pengganti B. Dalam hal formatnya, LIME dikenal sebagai metode penjelasan visual yang menggunakan model pengganti untuk menghasilkan penjelasan.

Penjelasan Aditif Shapley (SHAP)

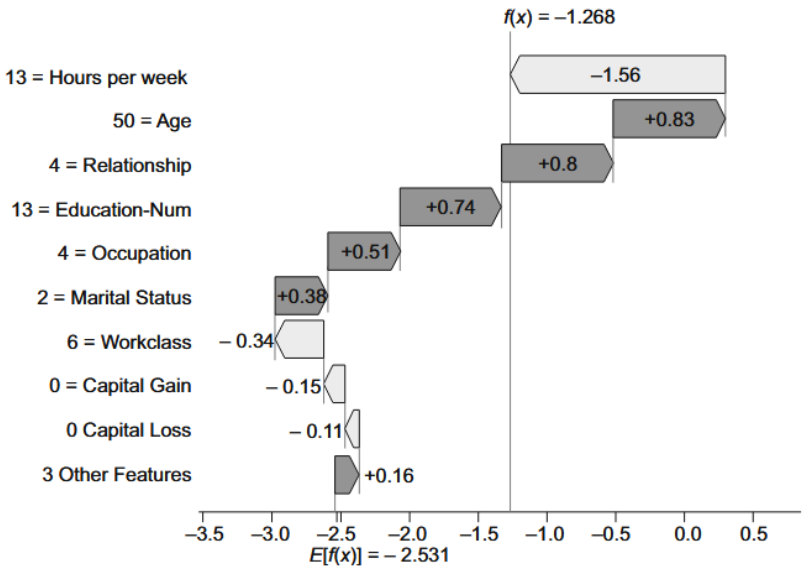
SHAP adalah metode XAI yang menetapkan nilai penting untuk setiap fitur untuk prediksi tertentu. SHAP diusulkan sebagai kerangka kerja yang didasarkan pada kelas ukuran penting fitur aditif, yang mencakup, misalnya, LIME. Kelas metode ini menggunakan model penjelasan yang merupakan fungsi linear dari variabel biner, yang masing-masing menunjukkan ada atau tidaknya fitur tertentu. SHAP menunjukkan bahwa LIME tidak hanya memiliki karakteristik ini, tetapi juga metode tradisional estimasi penting fitur yang dikenal sebagai Estimasi Nilai Shapley. Gagasan utama dari metode ini adalah bahwa efek dari tidak memasukkan fitur i bergantung pada fitur lain dalam model. Oleh karena itu, perbedaan hasil antara model yang menyertakan fitur i harus diukur terhadap semua kemungkinan subset fitur model yang tidak menyertakan i, dilambangkan sebagai $S \subset F \setminus \{i\}$.

Ini dihitung sebagai rata-rata tertimbang dari semua kemungkinan perbedaan:

$$\text{Value} = \sum_{S \subset F \setminus \{i\}.} \text{Weight}(S) \times (\text{Outcome}_{\text{with } i} - \text{Outcome}_{\text{without } i})$$

Formulasi ini memungkinkan pemahaman mendalam tentang bagaimana setiap fitur berkontribusi pada prediksi, dengan mempertimbangkan interaksi dengan semua fitur lainnya. Mengingat penjumlahan dalam persamaan tersebut mengandung $2|F|$, penjumlahan tersebut dapat didekati menggunakan pengambilan sampel atau dengan mengaproksimasikan efek penghapusan variabel i dengan mengintegrasikan sampel dari set pelatihan. Strategi aproksimasi apa pun yang digunakan, hal ini menghilangkan kebutuhan untuk melatih ulang model dan mengurangi jumlah suku yang perlu kita hitung dari $2|F|$. Strategi aproksimasi ini dikenal sebagai nilai pengambilan sampel Shapley, atau singkatnya nilai Shapley.

Dua properti krusial dalam model yang dapat dijelaskan: akurasi lokal dan konsistensi. Kita katakan bahwa model yang dapat dijelaskan g memenuhi akurasi lokal jika g cocok dengan keluaran f untuk keluaran asli x. Di sisi lain, konsistensi menyatakan bahwa jika suatu model berubah sehingga kontribusi masukan yang disederhanakan meningkat atau tetap tidak berubah terlepas dari masukan lain, atribusi masukan tersebut tidak boleh berkurang. Nilai Shapley adalah satu-satunya set nilai untuk model penting fitur aditif yang memenuhi akurasi lokal dan konsistensi.

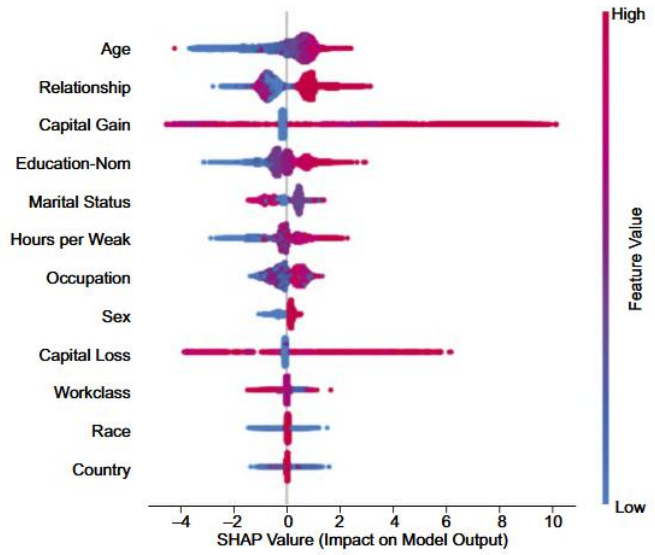


Gambar 7.5: Penjelasan SHAP lokal pada dataset pendapatan sensus dewasa standar dari repositori data pembelajaran mesin UCI.

SHAP menghasilkan penjelasan berdasarkan plot waterfall, yang memberikan penjelasan visual tentang pentingnya fitur untuk suatu contoh x. Kami akan mendemonstrasikan cara kerja SHAP menggunakan contoh dari dataset pendapatan sensus dewasa standar dari repositori data pembelajaran mesin UCI. Kami akan menghasilkan plot waterfall untuk satu observasi dan menghitung nilai Shapley untuk contoh tersebut. Hasilnya ditunjukkan pada Gambar 7.5.

Dalam plot air terjun, sumbu x merepresentasikan nilai variabel target, yang dalam hal ini menunjukkan apakah pendapatan melebihi Rp 831,7 juta per tahun. X adalah observasi, $f(x)$ adalah nilai yang diprediksi oleh model, dan $E[f(x)]$ adalah nilai ekspektasi pendapatan,

yaitu rata-rata dari semua prediksi yang dihasilkan. Nilai Shapley untuk setiap fitur dalam observasi menjelaskan seberapa besar fitur memengaruhi prediksi. Nilai ini menggambarkan deviasi prediksi dari nilai ekspektasi dan dengan demikian, menjelaskan kontribusi fitur tersebut dibandingkan dengan prediksi lainnya yang dibuat oleh model. Semakin besar nilai Shapley suatu fitur, semakin besar kontribusinya terhadap prediksi. Dalam contoh ini, plot air terjun menampilkan kontribusi dalam urutan menurun. Dapat juga diamati bahwa jumlah semua nilai Shapley berkorespondensi dengan $E[f(x)] - f(x)$.



Gambar 7.6: Penjelasan SHAP global pada dataset pendapatan sensus dewasa standar dari repositori data pembelajaran mesin UCI.

Dalam contoh ini, penjelasan mengilustrasikan efek dari dua subset fitur. Satu subset, yang terkait dengan jam kerja per minggu, kelas kerja, keuntungan modal, dan kerugian modal, cenderung memengaruhi prediksi dalam satu arah. Sebaliknya, fitur seperti usia, hubungan, pendidikan, pekerjaan, dan status perkawinan memengaruhi keputusan dalam arah yang berlawanan. Visualisasi ini juga menunjukkan bahwa terdapat tiga fitur lain dengan dampak yang lebih kecil, yang dikelompokkan ke dalam satu bin karena kami menggunakan maksimal sepuluh fitur dalam penjelasan.

SHAP juga memungkinkan pengilustrasian efek fitur pada keseluruhan dataset dengan melapiskan nilai Shapley dari setiap contoh. Hal ini dilakukan dalam plot kawanan lebah, seperti yang ditunjukkan pada Gambar 7.6. Dalam kasus ini, kami mengamati bahwa penjelasan yang dihasilkan oleh SHAP bersifat global, artinya penjelasan tersebut tidak bergantung pada contoh spesifik apa pun. Visualisasi ini membantu memahami dampak berbagai fitur terhadap keluaran model pembelajaran mesin. Setiap titik dalam plot mewakili nilai SHAP untuk suatu fitur dalam satu prediksi. Posisi pada sumbu X menunjukkan dampak fitur terhadap keluaran model, dengan nilai di sebelah kanan menunjukkan dampak yang lebih tinggi dan nilai di sebelah kiri menunjukkan dampak yang lebih rendah. Warna titik menunjukkan nilai fitur; merah menunjukkan nilai tinggi dan biru menunjukkan nilai rendah.

Fitur-fitur tersebut dicantumkan pada sumbu Y dan diurutkan berdasarkan dampak keseluruhannya terhadap model. Dalam plot, fitur seperti "Usia", "Hubungan", dan "Keuntungan Modal" memiliki dampak paling signifikan terhadap prediksi model.

Jenis visualisasi ini khususnya berguna untuk menginterpretasikan model kompleks dengan menunjukkan tidak hanya fitur mana yang penting, tetapi juga bagaimana fitur tersebut memengaruhi prediksi. Visualisasi ini membantu dalam memahami perilaku model dan dapat menjadi krusial untuk validasi model, debugging, dan menjelaskan prediksi kepada para pemangku kepentingan. Dari perspektif taksonomi Speith, SHAP diklasifikasikan sebagai metode XAI post-hoc yang bersifat model-agnostik. Metode ini beroperasi berdasarkan hasil untuk membangun penjelasan berdasarkan relevansi fitur. Dari sudut pandang fungsi, SHAP menggunakan mekanisme perturbasi, dan dalam hal cakupan, SHAP dapat beroperasi secara lokal maupun global.

BAB 8
TRANSFORMER DAN AI GENERATIF

8.1 PENDAHULUAN

Transformer merupakan landasan bagi banyak aplikasi AI paling signifikan saat ini. Dengan memanfaatkan arsitektur ini, sistem AI telah mampu menangani tugas-tugas kompleks seperti penerjemahan mesin, peringkasan teks, dan klasifikasi teks, tugas klasik dalam NLP. Selain berhasil mengelola tugas-tugas ini, arsitektur Transformer telah mendorong pengembangan AI Generatif, dengan aplikasi yang meluas hingga chatbot, dan penciptaan gambar serta video hiperrealistis. Kemajuan signifikan dalam AI Generatif dikaitkan dengan arsitektur ini dan elemen-elemen tambahan yang telah melengkapi pelatihan jaringan Transformer, seperti pembelajaran penguatan berdasarkan umpan balik manusia.

Seperti semua kemajuan teknologi utama, kemajuan AI Generatif menghadirkan tantangan etika yang sangat besar. Penggunaan teknologi ini secara tepat dapat membawa manfaat besar bagi masyarakat, tetapi penggunaan yang tidak tepat menimbulkan risiko yang harus dianalisis. Untuk mengatasi tantangan AI modern, kami akan mulai dengan menjelaskan bagaimana AI Generatif berfungsi untuk memahami potensinya. Selanjutnya, kami akan menelusuri studi kasus yang akan membantu mengilustrasikan tantangan etika yang melekat pada teknologi ini.

8.2 ARSITEKTUR TRANSFORMER

Perhatian adalah semua yang Anda butuhkan

Transformer adalah arsitektur jaringan saraf tiruan kompleks yang terdiri dari beberapa blok penyusun yang dikenal sebagai blok Transformer. Kita akan mulai dengan menjelaskan apa yang terkandung dalam blok Transformer dan kemudian mengeksplorasi bagaimana arsitektur ini diskalakan dengan memanfaatkan beberapa blok dasar. Inti dari arsitektur Transformer adalah mekanisme self-attention. Self-attention memungkinkan jaringan untuk mengekstrak informasi dari konteks dengan panjang yang bervariasi. Tidak seperti arsitektur pendahulunya, seperti jaringan rekursif, mekanisme self-attention memungkinkan Transformer untuk mengodekan dependensi di antara token dengan panjang input yang bervariasi, sehingga beroperasi pada dependensi jarak jauh di antara token input.

Mekanisme self-attention diimplementasikan menggunakan jaringan umpan-maju, dan operasi fundamentalnya serupa dengan yang ada di jaringan saraf tiruan konvensional lainnya, yaitu perkalian matriks-vektor. Misalkan kita memproses input teks melalui mekanisme self-attention. Dasar dari mekanisme ini melibatkan perbandingan item yang diminati (sebuah token) dengan token lain dalam input untuk mengungkapkan (dan mengodekan) relevansinya dalam konteks token yang sedang diproses. Bentuk perbandingan yang paling sederhana adalah perkalian titik.

Tanpa mengabaikan keumuman, mari kita asumsikan bahwa mekanisme self-attention mengikuti urutan pembacaan manusia (konteks kiri ke kanan). Ini berarti jika kita memproses

token ketiga dari input, misalnya x_3, kita akan mengodekan token ini ke dalam variabel baru, misalnya y_3. Karena Transformer dalam kasus ini mengikuti konteks kiri ke kanan, untuk menghitung y_3 kita harus mempertimbangkan tiga perkalian: x_3 dengan x_1, x_3 dengan x_2, dan x_3 dengan x_3. Perkalian ini dinormalisasi menggunakan lapisan softmax yang menghasilkan vektor bobot α_{ij}, yang menunjukkan relevansi token input j dengan token output i. Proses ini dinyatakan dalam persamaan berikut:

$$y_i = \text{softmax}\left(\sum_{j=1}^{N} \alpha_{ij} x_j\right)$$

di mana α_{ij} dihitung berdasarkan kesamaan antara token x_i dan x_j.

Mekanisme self-attention dasar ini menunjukkan bahwa bobot atensi α_{ij} dipelajari. Bobot ini dipelajari karena variabel x_1, x_2, dan x_3 dipelajari, karena jaringan tidak memproses token secara langsung, melainkan representasinya (pengodean). Karena x_1, x_2, dan x_3 adalah parameter, koefisien α_{ij} juga dipelajari. Kita katakan bahwa α_{ij} adalah koefisien atensi dari i ke token j.

Transformer meningkatkan mekanisme atensi dasar ini dengan mengodekan setiap token menurut tiga peran: kueri, kunci, dan nilai. Peran kueri menyiratkan bahwa pengodean masukan dianggap sebagai fokus atensi saat ini jika dibandingkan dengan masukan sebelumnya. Peran kunci menggunakan token sebagai masukan sebelumnya, sehingga penyisipan masukan dari langkah sebelumnya dibandingkan dengan yang berikutnya. Peran nilai menggunakan penyisipan masukan untuk menghitung nilai keluaran pada langkah waktu saat ini. Pada hakikatnya, untuk menghitung vektor peran ini, mekanisme perhatian-diri menggunakan tiga matriks parameter, yang menghitung q_i, k_i, dan v_i dari x_i berdasarkan produk matriks-vektor, seperti yang ditunjukkan dalam persamaan matriks berikut:

$$q_i = W^Q x_i, \;\; k_i = W^K x_i, \;\; v_i = W^V x_i$$

Di sini, W^Q, W^K, dan W^V masing-masing adalah matriks parameter untuk peran kueri, kunci, dan nilai.

Setelah kita memahami komponen kunci mekanisme ini, kita dapat memahami bagaimana y_i dihitung, yaitu, pengodean x_i yang dihitung menggunakan mekanisme self-attention. Pada dasarnya, alih-alih beroperasi langsung pada x_i, mekanisme self-attention beroperasi pada vektor kueri, kunci, dan nilai sebagai berikut:

$$y_i = \sum_{j \le i} \alpha_{ij} \, v_j,$$

$$\alpha_{ij} = q_i \cdot k_j$$

Persamaan-persamaan ini merepresentasikan bagaimana keluaran dihitung dengan membobotkan vektor nilai (v) dengan bobot atensi (α), yang ditentukan berdasarkan kesamaan antara kueri dan kunci.

Sebagaimana ditunjukkan oleh persamaan-persamaan ini, metode untuk menghasilkan pengkodean y_i melibatkan perbandingan kueri/kunci (produk) yang digunakan untuk menghitung koefisien atensi. Koefisien-koefisien ini kemudian dibandingkan dengan vektor nilai. Jumlah produk-produk ini menghasilkan vektor keluaran y_i. Mekanismenya dijelaskan dalam diagram yang ditunjukkan pada Gambar 8.1.

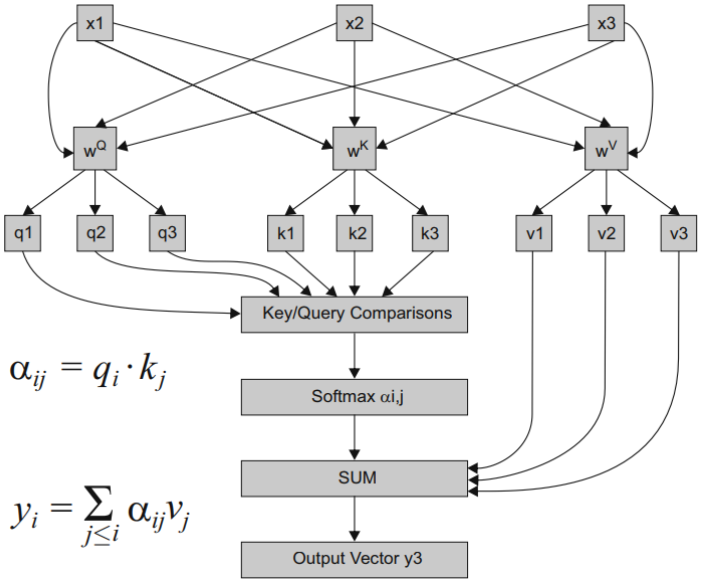
Aspek penting dari mekanisme ini adalah bahwa vektor keluaran dihitung berdasarkan kombinasi linear vektor nilai dari masukan, dengan koefisien atensi dihitung oleh mekanisme perbandingan kunci-kueri. Untuk memastikan bahwa kombinasi faktor-faktor ini memang linear, koefisien atensi dinormalisasi menggunakan lapisan softmax:

$$\text{Softmax} - \text{layer}(Q, K) = \text{logistic}\left(\frac{QK^T}{\sqrt{d_k}}\right)$$

di mana $\text{logistik}\,(z_i) = \frac{e^{z_i}}{\sum_j z_j}$ dan d_k adalah dimensionalitas vektor q dan k.

Langkah normalisasi ini memastikan bahwa koefisien berjumlah satu, memungkinkan kombinasi linear yang tepat dari vektor nilai untuk membentuk vektor keluaran. Fitur utama Transformer adalah ia tidak menghitung enkode secara berurutan tetapi dapat melakukannya secara paralel. Hal ini memungkinkan pemrosesan masukan yang efisien. Namun, ia memberlakukan batasan bahwa panjang masukan harus tetap, ditentukan oleh jumlah token masukan yang dapat diproses secara bersamaan melalui mekanisme self-attention. Ini adalah hiperparameter arsitektur jaringan.

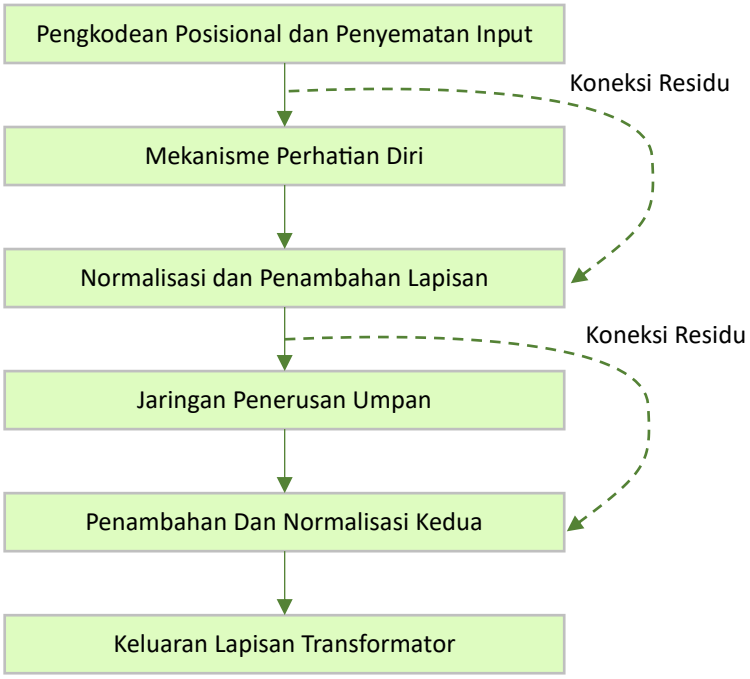
Relevansi mekanisme self-attention sangatlah substansial. Mekanisme ini memungkinkan Transformer untuk mengodekan dependensi jarak jauh antar simbol masukan. Secara spesifik, bergantung pada bagaimana Transformer dilatih, jaringan akan mempelajari embedding yang dirancang khusus untuk tugas yang dibutuhkan, dengan kapasitas signifikan untuk mengodekan dan memanfaatkan dependensi dengan panjang yang berubah-ubah pada masukan.



Gambar 8.1: Mekanisme self-attention dari transformer.

Sebuah blok transformer didasarkan pada mekanisme self-attention. Namun, blok ini juga mencakup operasi fundamental lainnya untuk mengodekan simbol masukan. Setelah melewati lapisan self-attention, koneksi residual diterapkan yang menggabungkan vektor keluaran dengan masukan. Koneksi residual ini didasarkan pada penjumlahan kedua vektor, yang kemudian dinormalisasi untuk memastikan bahwa penjumlahan beroperasi dalam domain terbatas. Terakhir, vektor dari lapisan Tambah dan Normalisasi diumpankan ke dalam lapisan umpan-maju. Baik pengodean masukan dari jaringan umpan-maju maupun keluarannya digabungkan kembali dalam koneksi residual kedua, menggunakan penjumlahan yang diikuti oleh normalisasi. Gambar 8.2 mengilustrasikan urutan operasi yang membentuk blok transformator.

Kita hampir selesai menjelaskan komponen-komponen dalam blok transformator. Elemen tambahan yang terintegrasi ke dalam arsitektur ini adalah atensi multi-head. Alih-alih beroperasi pada satu lapisan atensi mandiri, transformator memanfaatkan beberapa lapisan tersebut secara paralel, masing-masing dengan parameternya sendiri untuk menghitung vektor kueri, kunci, dan nilai. Keluaran dari lapisan paralel ini digabungkan menjadi satu vektor, yang kemudian dimasukkan ke dalam lapisan umpan-maju yang mengurangi dimensionalitas vektor. Tujuan mekanisme atensi multi-head adalah untuk meningkatkan skala kapabilitas transformator berdasarkan mekanisme atensi mandiri, dengan menambahkan lebih banyak parameter yang dapat dipelajari dan dengan demikian meningkatkan kapasitas pengodean transformator.



Gambar 8.2: Blok Transformer dimulai dengan penyandian posisi dan penanaman input, yang kemudian diteruskan ke mekanisme perhatian mandiri.

Komponen kunci lain dari arsitektur Transformer adalah penyandian posisi. Matriks parameter yang digunakan untuk menghitung vektor kueri, kunci, dan nilai dibagikan di berbagai token masukan. Akibatnya, transformer memiliki sifat ekuivalen permutasi terhadap urutan token masukan, yang berarti bahwa menukar urutan token tidak mengubah hasilnya.

Untuk mengodekan urutan token, vektor penyandian posisi dibangun untuk setiap posisi masukan. Penyandian ini digabungkan dengan penyandian token, membantu mempertahankan penyandian token dan posisinya secara bersamaan. Berbagai teknik digunakan untuk membangun penyandian posisi, dengan teknik berbasis fungsi sinusoidal menjadi yang paling umum digunakan.

Pengode-Dekoder Transformer

Blok-blok transformer dapat ditumpuk satu demi satu untuk menciptakan penyandian tingkat yang semakin tinggi. Jika masukan ke blok transformer pertama adalah serangkaian token data, kita katakan bahwa transformer beroperasi sebagai pengode transformator. Dimungkinkan untuk menempatkan lapisan linear yang diikuti oleh lapisan softmax di atas kosakata simbol pada keluaran blok terakhir dalam rantai transformator. Dalam kasus teks, lapisan softmax terakhir ini menyapu kosakata token. Apa yang dihasilkan transformator pada lapisan terakhir ini adalah probabilitas keluaran jaringan atas token-token tersebut. Kami menyebut strategi produksi simbol ini sebagai dekoder transformator.

Untuk membantu pembangkitan simbol, dekoder beroperasi menggunakan strategi umpan token yang disebut pembangkitan auto-regresif, di mana token masukan pada posisi i menghasilkan token keluaran pada posisi $i + 1$. Dengan menggeser pembangkitan token satu posisi, tugas yang ditangani oleh transformator disebut prediksi token berikutnya. Ini adalah dasar untuk pembangkitan teks dalam jaringan tipe transformator.

Kedua strategi operasi transformator, yaitu enkode dan dekode, dapat digunakan secara bersamaan. Ide kunci yang digunakan untuk menghubungkan kedua mode operasi ini adalah penambahan lapisan yang disebut perhatian encoder-decoder, yang menggunakan enkode dari encoder transformator dan menggabungkannya dengan enkode dari dekoder. Saat dekoder beroperasi secara autoregresif untuk prediksi token berikutnya, setiap simbol dari generator juga terhubung dengan enkode dari enkoder. Jenis pembangkitan bersyarat dengan atensi ini disebut atensi silang, karena dekoder transformator tidak hanya memperhatikan token masukan, tetapi juga enkode yang berasal dari enkoder transformator.

Strategi ini memungkinkan korelasi dua urutan token, urutan masukan enkoder dan urutan masukan dekoder, sebuah mode operasi pada arsitektur yang juga dikenal sebagai transformer seq2seq. Sementara urutan pertama beroperasi dalam mode enkoder untuk mengkondisikan dekoder, dekoder transformator beroperasi dalam mode auto-regresif untuk menghasilkan urutan token keluaran dari urutan token masukan dekoder. Metode pemberian makan ini merupakan dasar dari algoritma penerjemahan mesin, di mana kita menggunakan urutan token dalam enkoder dalam bahasa sumber dan menempatkan urutan token kedua dalam dekoder, dalam mode auto-regresif, dalam bahasa target.

Jika kita melakukan ini dengan banyak pasangan sekuens, yang melibatkan kumpulan data kalimat yang selaras, transformator akan belajar menghasilkan sekuens token dalam bahasa target dari sekuens dalam bahasa sumber. Dalam inferensi, ketika kita hanya memiliki teks dalam bahasa sumber, dekoder akan beroperasi berdasarkan pembangkitan regresi otomatis, dimulai dari simbol pertama sekuens, yang secara default selalu berupa simbol START.

Encoder Transformer

Arsitektur Transformer telah terbukti sangat berguna untuk menangani berbagai tugas dalam NLP. Salah satu implementasi penting melibatkan penggunaan encoder Transformer saja untuk tugas model bahasa bertopeng. Hal ini melibatkan penyambungan lapisan linier (umpan-maju) ke keluaran blok Transformer terakhir, diikuti oleh lapisan softmax untuk menjangkau kosakata token. Tugas model bahasa bertopeng memprediksi token keluaran dari urutan masukan yang beberapa tokennya telah ditopeng. Hal ini membentuk fondasi Representasi Encoder Dua Arah dari Transformer (BERT), yang membangun penyandian kata dengan melatih Transformer pada volume teks besar menggunakan tugas model bahasa bertopeng.

BERT dilatih pada korpus teks yang beragam dan luas, termasuk BookCorpus, yang berisi lebih dari 800 juta kata, dan Wikipedia, dengan total 2,5 miliar kata. Pemanfaatan volume teks yang besar memberikan model-model ini kemampuan generalisasi yang lebih baik. Namun, penting untuk dicatat bahwa model ini akan mereplikasi bias yang terdapat dalam teks-teks ini, yang tidak diproses sebelumnya untuk mengatasi bias tersebut. Model memproses teks sebagaimana adanya.

Setelah encoder Transformer dilatih menggunakan strategi BERT, kita dapat memasukkan kalimat baru selama inferensi. Dengan melakukan forward pass melalui jaringan, kita dapat mengambil enkode kata dari blok Transformer, menggunakan vektor-vektor ini

sebagai penyisipan kata yang bergantung pada konteks. Penyisipan kata sangat berguna karena membantu kita membangun representasi vektor dokumen dari penyisipan kata dengan menggabungkan vektor-vektor tersebut menggunakan operasi agregasi seperti perata-rataan, sebuah strategi yang dikenal sebagai penyisipan kata rata-rata. Berdasarkan vektor-vektor ini, kita dapat melatih pengklasifikasi dokumen. Ada juga cara lain untuk memanfaatkan vektor BERT, yaitu dengan menggunakan strategi agregasi tingkat kalimat yang dikenal sebagai Sentence BERT, yang menyediakan penyisipan kalimat untuk tugas-tugas tingkat kalimat.

BERT juga telah digunakan sebagai model yang telah dilatih sebelumnya. Setelah encoder Transformer dilatih menggunakan BERT, kita dapat menyempurnakan jaringan ini untuk tugas-tugas hilir dengan memberikan masukan dan keluaran dari set data yang dianotasi secara khusus. Pendekatan ini, yang dikenal sebagai fine-tuning, telah berhasil mengevaluasi BERT dalam berbagai tugas NLP, termasuk klasifikasi dokumen, analisis sentimen, dan tanya jawab tertutup.

Meskipun penggunaan penyisipan kata menawarkan manfaat yang signifikan, hal ini juga menghadirkan beberapa risiko. Model pra-latihan seperti BERT bergantung pada volume teks besar yang mengandung bias. Kita akan melihat bahwa analisis berdasarkan penyisipan kata menggambarkan efek bias ini terhadap representasi yang dipelajari. Strategi kita akan menunjukkan bahwa analogi yang diambil dari penyisipan kata mengungkapkan reproduksi bias budaya dan historis.

Dekoder Transformer

Sebagaimana encoder transformer telah digunakan untuk membangun representasi, dekoder transformer telah dimanfaatkan untuk pembangkitan. Dekoder transformer membentuk dasar dari apa yang kita kenal sebagai AI Generatif. Dengan mengandalkan mekanisme pembangkitan auto-regresif, dekoder transformer dapat digunakan untuk melatih model generatif. Kemajuan terbaru dalam AI generatif sebagian besar disebabkan oleh adopsi arsitektur ini.

Dekoder transformator dapat beroperasi tanpa enkoder dalam mode pembangkitan teks regresif otomatis. Berdasarkan tugas prediksi token berikutnya, dekoder transformator menggunakan varian model bahasa bertopeng yang disebut model bahasa kausal. Strategi di sini melibatkan modifikasi lapisan perhatian-diri agar hanya berfokus pada token sebelumnya dalam urutan masukan. Tidak seperti model bahasa bertopeng, yang mempertimbangkan konteks di sebelah kiri dan kanan simbol, oleh karena itu BERT merupakan transformator dua arah, model bahasa kausal hanya memperhatikan token sebelumnya. Dengan memaksa mekanisme perhatian untuk bertindak kausal, dekoder siap untuk menghasilkan teks menggunakan mekanisme ini yang dikombinasikan dengan strategi pembangkitan regresif otomatis.

Dekoder transformator dapat menggunakan beberapa blok transformator. Serupa dengan enkoder, dekoder transformator menggunakan penempatan posisional pada masukan dan softmax pada keluaran. Lapisan terakhir ini menghasilkan probabilitas keluaran atas kosakata token. Ketika dilatih pada tugas model bahasa kausal, dekoder belajar menghasilkan

teks berdasarkan prediksi token berikutnya. Tugas pelengkapan teks ini merupakan fondasi LLM.

Dekoder transformator yang dilatih untuk melengkapi teks pada volume teks yang besar dapat menghasilkan teks berdasarkan teks masukan. Seperti BERT, LLM awal yang berbasis dekoder transformator dilatih menggunakan dataset BooksCorpus. Hal ini mengarah pada pengembangan GPT-1, model pertama dari OpenAI yang dikenal dengan akronimnya, Generative Pretrained Transformer. Makalah asli tentang GPT-1 menggunakan fine-tuning pada model dasar untuk menilai kapabilitas GPT-1 dalam tugas-tugas hilir. Seperti halnya BERT, fine-tuning transformator yang telah dilatih sebelumnya secara signifikan meningkatkan berbagai tugas NLP, termasuk tanya jawab tertutup, kesamaan semantik, dan klasifikasi teks.

8.3 UMPAN BALIK MANUSIA UNTUK TRANSFORMER

Arsitektur transformer merupakan inovasi disruptif dalam AI. Meskipun mekanisme self-attention dan peluang yang diberikannya untuk melatih model menggunakan strategi tanpa pengawasan merupakan fitur-fitur penting, hubungannya dengan pembelajaran penguatan kemungkinan besar telah menarik perhatian paling besar dalam beberapa tahun terakhir. Secara spesifik, hubungan antara arsitektur Transformer dan Pembelajaran Penguatan Berbasis Umpan Balik Manusia (RLHF) merupakan perkembangan signifikan di bidang AI, khususnya dalam NLP.

Pengumpulan umpan balik manusia telah terbukti sangat berharga dalam bekerja dengan LLM, terutama untuk tugas-tugas yang melibatkan teks seperti pembuatan ringkasan. Stiennon dkk. menggabungkan umpan balik manusia pada unggahan Reddit dengan menerapkan berbagai strategi otomatis untuk menghasilkan ringkasan untuk unggahan tertentu. Setelah daftar kandidat ringkasan tersedia, dua di antaranya dipilih untuk dievaluasi oleh manusia. Anotator manusia kemudian menilai ringkasan mana yang lebih mewakili unggahan asli, yang memungkinkan terciptanya model penghargaan. Untuk proses ini, dua ringkasan dari satu unggahan dimasukkan ke dalam model penghargaan, yang kemudian menghitung penghargaan untuk setiap ringkasan. Kerugian model ditentukan berdasarkan imbalan dan label manusia ini, yang kemudian digunakan untuk memperbarui model imbalan.

Kami akan menggunakan model imbalan untuk mengawasi penyempurnaan transformator pada tugas peringkasan teks. Proses ini melibatkan pengambilan sampel postingan baru dari kumpulan data dan memberikan ringkasan yang dihasilkan oleh model kami. Model imbalan kemudian menetapkan imbalan untuk ringkasan ini, yang digunakan untuk menyempurnakan model kami lebih lanjut. Prinsip penggunaan model imbalan untuk penyelarasan model berdasarkan umpan balik manusia mendorong pengembangan Instruct GPT. Instruct GPT adalah LLM yang dirancang untuk menghasilkan keluaran yang dikondisikan pada prompt, yang dapat mencakup instruksi, informasi kontekstual, atau contoh tugas yang harus diselesaikan. Instruct GPT menggunakan GPT-3 sebagai model dasarnya, versi GPT-2 yang disempurnakan dengan parameter yang jauh lebih banyak dan dilatih pada korpus teks yang lebih besar.

Instruct GPT diselaraskan menggunakan kumpulan data prompt yang mencakup berbagai tugas. Ketika prompt dari kumpulan data ini diambil sampelnya, pelabel manusia menunjukkan perilaku keluaran yang diinginkan dengan menuliskan respons yang diharapkan. Data ini kemudian digunakan untuk menyelaraskan GPT-3. Pada fase selanjutnya, prompt dan beberapa keluaran model diambil sampelnya, dan pelabel manusia memeringkat keluaran ini dari terbaik hingga terburuk. Data ini melatih model reward. Setelah model reward dilatih, model tersebut mengoptimalkan model berdasarkan reward yang telah dihitung. Prompt baru diambil sampelnya, model menghasilkan output, dan model reward mengevaluasi output ini. Reward yang dihasilkan digunakan untuk memperbarui model.

Instruct GPT mencakup peningkatan seperti rentang tugas yang lebih luas dalam dataset prompt-nya, tidak terbatas pada ringkasan. Tugas-tugas ini meliputi pembuatan teks terbuka, tanya jawab terbuka, curah pendapat, obrolan, parafrase, peringkasan, klasifikasi, tanya jawab tertutup (pilihan ganda), dan ekstraksi teks. Meskipun model dasar mencapai hasil yang menarik pada tugas-tugas ini, yang menunjukkan kemampuan pemrosesan teks model, model yang diselaraskan dengan dataset prompt menunjukkan hasil yang jauh lebih baik. Peningkatan ini sangat bergantung pada penyelarasan keluaran model dengan prompt.

Instruct GPT membentuk dasar ChatGPT, sebuah chatbot OpenAI yang impresif dan mampu memproses perintah serta menghasilkan respons yang selaras. ChatGPT dibangun di atas GPT3.5, sebuah model dasar yang memperluas kapabilitas GPT-3. Namun, OpenAI telah memperbarui model dasarnya, merilis GPT-4 pada tahun 2023, LLM tercanggih mereka hingga saat ini. Kapabilitas ChatGPT berbasis GPT-4 begitu luar biasa sehingga digambarkan sedang bergerak menuju AI umum, yaitu AI yang mampu menangani tugas-tugas baru yang tidak dilatih secara khusus. Implikasi dan risiko etis dari teknologi ini akan menjadi fokus utama bab selanjutnya.

8.4 KESIMPULAN

Arsitektur Transformer, khususnya dalam bentuk ChatGPT, telah memberikan pengaruh yang mendalam pada berbagai sektor, termasuk pendidikan, layanan kesehatan, dan layanan pelanggan. ChatGPT telah menarik perhatian signifikan di bidang pendidikan karena potensinya untuk membentuk kembali norma-norma pendidikan dan menawarkan pengalaman belajar yang personal dan adaptif. Namun, terdapat kekhawatiran mengenai potensinya untuk mengurangi keterampilan analitis dan mendorong perilaku buruk. Dalam layanan kesehatan, iterasi terbaru, GPT-4, dikenal karena basis pengetahuannya yang luas dan kemampuan pemecahan masalah yang ditingkatkan. Khususnya, kemampuan barunya untuk menganalisis gambar menjanjikan untuk interpretasi dan diagnosis citra medis.

Selain itu, ChatGPT telah dinilai kemampuannya untuk merespons pertanyaan klinis yang kompleks, menunjukkan potensi sebagai alat interaktif untuk pendidikan kedokteran. Dalam layanan pelanggan, ChatGPT telah digunakan untuk menyediakan logika dan konteks informasional di sebagian besar respons, yang menunjukkan penerapannya di bidang ini. Meskipun demikian, terdapat kekhawatiran etis terkait penggunaan ChatGPT, terutama terkait potensi penyebaran misinformasi, isu privasi, dan risiko ketergantungan yang berlebihan pada

teknologi. Penelitian di masa mendatang diperlukan untuk mengatasi keterbatasan dan potensi risiko ini, dengan penekanan pada pertimbangan dan verifikasi informasi yang diberikan secara cermat.

Singkatnya, meskipun arsitektur Transformer, sebagaimana dicontohkan oleh ChatGPT, berpotensi merevolusi berbagai industri, penting untuk mempertimbangkan keterbatasan dan implikasi etisnya secara cermat guna memastikan integrasinya yang bertanggung jawab ke dalam domain-domain tersebut.

BAB 9
NLP DAN BIAS REPRESENTASIONAL

9.1 PENDAHULUAN

Dalam NLP, bias telah menjadi tantangan yang terus-menerus, dengan akarnya sering kali ditelusuri kembali ke representasi yang dihasilkan oleh penyisipan kata. Penyisipan kata ini, yang berfungsi sebagai representasi berbasis vektor dari kata-kata dalam ruang semantik, rentan terhadap bias karena mencerminkan data tekstual asalnya. Bias, dalam konteks ini, biasanya didefinisikan oleh asosiasi negatif yang muncul dalam representasi kata ini, yang memengaruhi bagaimana kata-kata diposisikan relatif satu sama lain. Melalui penggunaan analogi, seperti ungkapan terkenal "Pria bagi komputer seperti wanita bagi ibu rumah tangga," para peneliti telah mengilustrasikan tingkat bias yang ada dalam penyisipan kata dan mengevaluasi potensi dampak hilirnya pada berbagai tugas NLP menggunakan analogi. Keberadaan bias dalam representasi ini diilustrasikan, dan potensi dampaknya pada tugas-tugas hilir dikaji.

Sebagaimana dibahas dalam bab sebelumnya, arsitektur Transformer saat ini dominan. Namun, sebelum Transformer, teknik-teknik yang lazim dalam NLP untuk pembelajaran representasi terutama bergantung pada penyisipan kata yang independen terhadap konteks. Vektor-vektor ini dihitung baik melalui jaringan umpan-maju (word2vec) atau melalui strategi faktorisasi matriks dari matriks term-dokumen dari korpus teks (GloVe). Vektor-vektor ini merepresentasikan posisi sebuah kata dalam ruang representasi semantik. Kedekatan antara dua kata menunjukkan hubungan semantik, sementara jarak menunjukkan hubungan semantik yang lemah. Dalam kedua strategi tersebut, vektor diambil dari model yang telah dilatih sebelumnya pada sejumlah besar teks. Penggunaan korpus untuk melatih model-model ini didasarkan pada prinsip kedekatan tekstual: jika dua kata cenderung muncul bersamaan dalam teks, maka kata-kata ini terhubung. Dengan menggunakan teks untuk membangun hubungan semantik antar kata, penyisipan kata mereproduksi bias dan stereotip karena teks menggambarkan dunia sebagaimana adanya, sehingga mereproduksi bias yang sudah ada.

Analisis bias dalam NLP telah didekati dari berbagai perspektif, misalnya, strategi untuk mengkuantifikasi bias berdasarkan metrik telah ditetapkan. Demikian pula, strategi mitigasi bias telah diusulkan, yang sebagian besar merupakan strategi post-hoc yang melakukan penyesuaian pada penempatan kata untuk mencapai tujuan tertentu. Sebagian besar studi ini berfokus pada analisis berdasarkan kelompok yang kurang beruntung, dengan penekanan kuat pada analisis gender. Dalam hal atribut yang menarik, reproduksi stereotip dalam profesi menonjol.

Terdapat pula minat yang semakin besar untuk mempelajari efek kontrafaktual dalam teks, yang diusulkan sebagai strategi mitigasi dan analisis korpus yang memungkinkan pengkajian teks dengan lebih sedikit bias dan stereotip. Kami akan mulai dengan mengilustrasikan isu-isu di area ini berdasarkan sebuah karya perintis dalam analisis bias dalam NLP: Pria bagi komputer sebagaimana wanita bagi ibu rumah tangga.

9.2 ANALOGI DAN STEREOTIP KATA

Debias keras

menunjukkan bahwa penyisipan kata yang dilatih pada korpus yang digunakan untuk membangun representasi kata, seperti GloVe, menunjukkan stereotip gender untuk pria dan wanita. Mereka memperkenalkan konsep kunci berdasarkan ide geometris, yang bertujuan untuk representasi ideal yang menjadi dasar aspirasi model yang bebas bias. Konsep ini bergantung pada pembedaan antara kata-kata netral gender dan kata-kata yang bias gender, dengan mencatat bahwa kata-kata netral gender dapat dipisahkan secara linear dari kata-kata definisi gender dalam ruang penyisipan kata. Berdasarkan sifat-sifat ini, mereka mengusulkan metodologi untuk memodifikasi penyisipan dengan menghilangkan stereotip gender, mengeliminasi asosiasi stereotip sekaligus mempertahankan asosiasi deskriptif antar kata.

Analisis dimulai dengan mendefinisikan apa itu analogi kata. Dengan tiga kata, misalnya, "dia," "dia," dan "raja," kami mencari kata keempat untuk melengkapi analogi: "dia bagi raja sebagaimana dia bagi x." Penyisipan kata seperti GloVe atau Word2Vec menemukan analogi yang mencengangkan. Meskipun dalam contoh ini, sebagian besar model yang telah dilatih sebelumnya akan menyelesaikan x sebagai "ratu", dalam analogi lain yang melibatkan profesi, hasilnya berbeda secara signifikan. Misalnya, analogi "pria bagi dokter seperti wanita bagi x" umumnya diselesaikan sebagai "perawat". Demikian pula, "pria bagi pemrogram komputer seperti wanita bagi x" biasanya menghasilkan "ibu rumah tangga". Analogi-analogi ini jelas mereproduksi stereotip gender.

memperkenalkan metode yang disebut hard-debias, yang didasarkan pada identifikasi subruang gender. Metode ini dimulai dengan sekumpulan kata untuk dinetralkan dan bekerja dengan penyisipan kata dari sekumpulan kata, yang bersifat deskriptif gender. Untuk setiap sekumpulan kata, rata-rata penyisipan kata yang menyusunnya didefinisikan, dilambangkan sebagai ui. Kemudian, subruang bias *B* didefinisikan sebagai k baris pertama Dekomposisi Nilai Singular (SVD) dari matriks vektor deviasi di sekitar rata-rata untuk semua penyisipan kata yang membentuk sekumpulan kata tersebut. Jika himpunan kata menggambarkan kata gender, misalnya, memiliki dua himpunan kata, satu untuk laki-laki dan satu untuk perempuan, B akan memungkinkan identifikasi subruang gender dari korpus yang sedang kita kerjakan.

Setelah menentukan subruang gender, pekerjaan dilakukan pada kata-kata yang akan dinetralkan, seperti kata-kata profesional. Untuk setiap kata yang akan dinetralkan, penyisipan baru, yang penulis sebut penyisipan ulang, dilambangkan sebagai w, mengoreksi penyisipan kata asli dengan vektor rata-rata subruang gender, yang kita nyatakan sebagai $\vec{w}B$. Kemudian, untuk setiap kata dalam himpunan kata yang akan dinetralkan, kita menghitung rata-rata atas grup berdasarkan vektor penyisipan ulang, dilambangkan sebagai μ. Selanjutnya, vektor yang mengukur selisih antara rata-rata grup dan vektor rata-rata subruang gender dihitung, yaitu, $v = \mu - \mu_B$, di mana μ_B merepresentasikan vektor rata-rata B. Akhirnya, untuk setiap kata yang akan dinetralkan, penyisipan kata dihitung sebagai:

$$\vec{w} = v + \sqrt{1 - \|v\|^2} \frac{\vec{w}_B - \mu_B}{\|\vec{w}_B - \mu_B\|}$$

Efek dari metode Hard Debias didasarkan pada penol-an proyeksi gender setiap kata sepanjang arah gender yang telah ditentukan sebelumnya. Metode penghitungan ulang embedding ini mengasumsikan bahwa kondisi yang adil untuk representasi didasarkan pada simetri kata-kata yang sensitif gender, seperti profesi, di sekitar sumbu gender. Ini adalah ide geometris yang sederhana namun penting. Ruang representasi dikatakan adil jika kata-kata yang akan dinetralkan, seperti profesi, bersifat netral gender. Ini berarti bahwa, dalam subruang gender, kata-kata yang akan dinetralkan harus sangat berdekatan satu sama lain, untuk mencegah refleksi stereotip gender dalam profesi.

Konsep yang diturunkan dari ide Bolukbasi ini memerlukan definisi implisit tentang apa yang dianggap netral gender. Menurut pendekatan Hard Debias, tidak ada bias gender jika setiap kata dalam kosakata yang tidak secara eksplisit bergender berjarak sama dari kedua elemen dari semua pasangan kata yang secara eksplisit bergender. Akibatnya, keadilan dalam ruang representasi serupa dengan konsep simetri.

Lipstik pada babi: Metode debias tidak menghilangkannya

Gonen & Goldberg menganalisis metode debias keras dan pendekatan-pendekatan selanjutnya, yang semuanya didasarkan pada konsep simetri dan netralitas terkait arah gender dalam ruang representasi. Melalui eksperimen pengelompokan dalam ruang penempatan kata, mereka mengungkapkan bahwa debias keras tidak mampu menetralkan asosiasi antarkata akibat bias implisit residual dalam komponen representasi yang tidak selaras dengan subruang gender. Masalah ini muncul karena hubungan kedekatan yang mengkodekan bias gender tidak terbatas pada subruang gender, tetapi tetap ada di luarnya.

Persistensi ini disebabkan oleh bias proyeksi yang diandalkan oleh debias keras, yang berkorelasi dengan bias tetangga, sehingga mempertahankan hubungan kedekatan ruang asli berdasarkan struktur lingkungan. Pengelompokan kata gender pada penempatan kata yang telah didebaskan keras dari kata-kata yang berjenis kelamin tidak secara efektif mengungkapkan bias tersebut. Misalnya, "perawat" tidak lagi dekat dengan kata-kata yang ditandai secara eksplisit sebagai kata feminin. Namun, bias tersebut masih muncul ketika kata tersebut berdekatan dengan kata-kata feminin yang ditandai secara sosial, seperti "penata rambut" atau "kapten". Hal ini menunjukkan bahwa cara yang efektif untuk mengukur bias tidak hanya berdasarkan simetri terkait arah gender, tetapi juga persentase kata-kata yang bias secara sosial pria/wanita di antara k tetangga terdekat kata target.

Metafora "lipstik pada babi", yang menggambarkan bahwa metode debias berdasarkan simetri hanya menghilangkan bias secara dangkal, menunjukkan bahwa kata-kata dengan bias gender yang kuat mudah dikelompokkan bersama. Dengan demikian, asosiasi antarkata stereotip dapat bertahan dalam metode debias berdasarkan simetri karena metode tersebut mempertahankan asosiasi struktural ruang representasi berdasarkan lingkungan. Hal ini juga menggambarkan bahwa kata-kata dengan gender implisit dari stereotip sosial (misalnya, "penata rambut" atau "kapten") masih cenderung mengelompok dengan kata-kata gender implisit lain dengan gender yang sama, serupa dengan penempatan kata yang tidak didebias. Dengan demikian, gender implisit kata-kata dengan bias sebelumnya yang lazim mudah diprediksi hanya berdasarkan vektornya saja.

Keterbatasan metode berbasis subruang netralitas gender

Pembaca mungkin menyadari beberapa keterbatasan metode debias berdasarkan konsep simetri. Pertama, banyak dari metode ini, terutama Hard DeBias yang menjadi pionir, bergantung pada definisi himpunan kata. Pemilihan himpunan kata dan kata yang akan dinetralkan merupakan proses yang rumit, karena bias dapat muncul secara tidak sengaja selama pemilihan ini. Intinya, hal ini menimbulkan pertanyaan tentang siapa yang memutuskan kata mana yang harus dinetralkan dan kata mana yang mewakili poros masyarakat yang perlu dilindungi. Definisi semacam itu secara inheren mengandung semua risiko yang terkait dengan definisi buatan manusia, mereproduksi bias dan stereotip melalui bias seleksi.

Keterbatasan kedua adalah bahwa metode ini didasarkan pada premis bahwa penyisipan kata bersifat independen terhadap konteks. Ini menyiratkan bahwa representasi suatu kata bersifat statis, dan oleh karena itu, representasi tersebut tetap sama terlepas dari konteks penggunaannya. Metode seperti Word2vec dan GloVe beroperasi berdasarkan prinsip ini. Masalah dengan penyisipan kata yang independen terhadap konteks adalah bahwa hal tersebut tidak memperhitungkan polisemi, sebuah kata dapat memiliki banyak arti, dan arti yang tepat bergantung pada kalimat di mana kata tersebut digunakan.

Oleh karena itu, dengan mengabaikan ketergantungan konteks, kita menghambat kemampuan untuk menangkap representasi kata yang tepat dalam konteks penggunaannya, sehingga mengabaikan efek polisemi. Oleh karena itu, representasi kata yang bergantung pada konteks akan lebih baik untuk mengatasi polisemi secara efektif. Inilah yang dicapai BERT, karena memungkinkan pengambilan vektor yang dikondisikan pada kalimat di mana kata tersebut digunakan.

9.3 AUGMENTASI DATA KONTRAFAKTUAL

Meskipun metode yang didasarkan pada simetri dan koreksi penempatan kata pada dasarnya bersifat post-hoc artinya metode tersebut beroperasi pada representasi yang telah dihitung sebelumnya dan kemudian melakukan intervensi untuk melakukan koreksi—pengolahan data sebelum digunakan untuk melatih model juga dimungkinkan. Pendekatan ini didasarkan pada asumsi bahwa data, sebagaimana adanya, mereproduksi bias dan stereotip. Dalam hal ini, data mentah merepresentasikan dunia sebagaimana adanya, lengkap dengan segala asimetri dan biasnya. Oleh karena itu, cara yang tepat untuk mengurangi bias adalah dengan memodifikasi data agar merepresentasikan dunia sebagaimana mestinya.

Augmentasi Data Kontrafaktual (CDA) bertujuan untuk menciptakan versi alternatif data yang mengurangi bias inheren dalam data. Gagasan mendasar dari strategi ini adalah bahwa dengan mendeteksi penyebutan stereotip dalam sebuah kalimat, kita dapat mengurangi efeknya yang berpotensi merugikan dengan menyediakan versi alternatif pada kumpulan data yang menyeimbangkan penyebutan di berbagai kelompok. Misalnya, perhatikan kalimat sumber "Seorang pria sedang berjalan." Karena "pria" adalah kata yang termasuk dalam kelompok kata yang berorientasi pada laki-laki, kita dapat menyeimbangkan efek kalimat ini untuk mencegah tindakan berjalan semata-mata dikaitkan dengan laki-laki.

Dengan mengganti "pria" dengan "wanita", kita menghasilkan kalimat baru, "Seorang wanita sedang berjalan", yang menyeimbangkan efek gender dalam data.

Salah satu keuntungan dari teknik augmentasi data ini adalah sifatnya yang agnostik terhadap model, artinya teknik ini dapat beroperasi pada model penyandian teks apa pun, termasuk model yang mempertimbangkan konteks kata seperti transformer. Inilah yang dilakukan, menggunakan berbagai strategi augmentasi data untuk memodifikasi set data pelatihan untuk BERT. Awalnya, penulis menunjukkan bahwa BERT dan versi sulingnya, ALBERT, mempelajari dan memanfaatkan korelasi gender.

Oleh karena itu, mereka menerapkan augmentasi data kontrafaktual pada penyebutan gender untuk mengurangi korelasi gender. Untuk menerapkan pra-pelatihan kontrafaktual pada BERT, penulis menghasilkan contoh pelatihan tambahan dari Wikipedia bahasa Inggris menggunakan pasangan kata gender (misalnya, dia – dia). Pertama, mereka mengidentifikasi kalimat yang mengandung salah satu kata yang bergender, lalu menghasilkan kalimat kontrafaktual dengan mengganti pasangan gender kata tersebut. Hasil BERT dan ALBERT menunjukkan bahwa korelasi berbasis gender menurun menggunakan teknik ini. Temuan penting lainnya dari studi ini menyoroti bahwa model-model ini tangguh terhadap fine-tuning; artinya, setelah mengurangi korelasi berbasis gender dan menggunakan model yang telah dilatih sebelumnya untuk tugas baru, korelasi gendernya tetap rendah. Aspek menarik dari strategi ini adalah keakuratan model yang telah dilatih sebelumnya tetap terjaga saat digunakan dalam tugas-tugas selanjutnya. Hal ini menunjukkan bahwa kita dapat mengurangi bias dalam model yang telah dilatih sebelumnya tanpa memperhitungkan biaya terkait utilitas model.

9.4 KETERBATASAN STRATEGI DEBIAS MODEL

Meskipun metode yang didasarkan pada simetri dan koreksi penempatan kata secara inheren bersifat post-hoc, artinya metode tersebut beroperasi pada representasi yang telah dihitung sebelumnya dan kemudian melakukan intervensi untuk melakukan koreksi, pengolahan data sebelum digunakan untuk melatih model juga dimungkinkan. Pendekatan ini didasarkan pada asumsi bahwa data, sebagaimana adanya, mereproduksi bias dan stereotip. Dengan demikian, data mentah merepresentasikan dunia sebagaimana adanya, dengan segala asimetri dan biasnya. Oleh karena itu, cara yang tepat untuk mengurangi bias adalah dengan memodifikasi data agar merepresentasikan dunia sebagaimana mestinya. Augmentasi Data Kontrafaktual (CDA) bertujuan untuk menciptakan versi alternatif dari data yang mengurangi bias inheren. Gagasan mendasar dari strategi ini adalah bahwa dengan mendeteksi penyebutan stereotip dalam sebuah kalimat, kita dapat menangkal potensi efek buruknya dengan menyediakan versi alternatif pada kumpulan data yang menyeimbangkan penyebutan di antara kelompok-kelompok lain. Misalnya, perhatikan kalimat sumber "Seorang pria sedang berjalan". Karena "pria" termasuk dalam himpunan kata yang terdiri dari istilah laki-laki, kita dapat menyeimbangkan efek kalimat ini untuk mencegah tindakan berjalan semata-mata dikaitkan dengan laki-laki. Dengan mengganti "pria" dengan "wanita", kita memperoleh

kalimat baru, "Seorang wanita sedang berjalan", yang menyeimbangkan efek dalam data penyebutan yang hanya berorientasi pada satu kelompok gender.

Keunggulan teknik augmentasi data ini adalah sifatnya yang agnostik terhadap model, sehingga dapat beroperasi pada model pengodean teks apa pun, termasuk model yang sadar konteks seperti transformer. Inilah yang dilakukan, yang menggunakan berbagai strategi augmentasi data untuk memodifikasi himpunan data pelatihan untuk BERT. Pertama, penulis menunjukkan bahwa BERT dan versi sulingnya, ALBERT, mempelajari dan memanfaatkan korelasi gender.

Oleh karena itu, mereka menggunakan augmentasi data kontrafaktual pada penyebutan gender untuk mengurangi korelasi gender. Untuk menerapkan pra-pelatihan kontrafaktual pada BERT, penulis menghasilkan contoh pelatihan tambahan dari Wikipedia bahasa Inggris menggunakan pasangan kata gender (misalnya, dia – dia). Mereka pertama-tama menemukan kalimat yang mengandung salah satu kata yang bergender, kemudian menghasilkan kalimat kontrafaktual dengan mengganti pasangan gender kata tersebut. Hasil BERT dan ALBERT menunjukkan bahwa korelasi berbasis gender menurun menggunakan teknik ini.

Temuan penting lainnya dari studi ini menyoroti bahwa model-model ini tangguh terhadap fine-tuning, artinya setelah mengurangi korelasi berbasis gender dan menggunakan model yang telah dilatih sebelumnya untuk tugas baru, korelasi berbasis gender tetap rendah. Aspek menarik lainnya dari strategi ini adalah mempertahankan akurasi model yang telah dilatih sebelumnya saat digunakan dalam tugas-tugas selanjutnya. Hal ini menggarisbawahi kemungkinan pengurangan bias dalam model yang telah dilatih sebelumnya tanpa menimbulkan biaya terkait utilitas model.

9.5 KESIMPULAN

Metode debiasing dalam NLP secara tradisional berfokus pada isolasi atau penghapusan informasi terkait atribut sensitif. Namun, terdapat argumen yang semakin kuat untuk pemanfaatan informasi sensitif ini secara 'adil', yang didukung oleh penjelasan, alih-alih penghapusannya secara menyeluruh. Perspektif ini menunjukkan bahwa atribut sensitif harus digunakan secara bijaksana, dengan penekanan pada transparansi dan keadilan dalam pemrosesan data. Selain itu, integrasi pengaturan interaktif yang menyertakan umpan balik pengguna telah diusulkan sebagai cara untuk mencapai pendekatan yang lebih seimbang dan adil dalam mitigasi bias. Metode ini tidak hanya meningkatkan kinerja tugas tetapi juga meningkatkan pengurangan bias dalam penjelasan yang diberikan oleh model, sekaligus mempertahankan akurasi prediksi.

Lebih lanjut, strategi yang independen dari model spesifik, yang dikenal sebagai strategi debiasing model-agnostik, telah dikembangkan untuk memperkuat model NLP terhadap berbagai serangan adversarial. Strategi ini dirancang untuk mempertahankan atau meningkatkan kemampuan generalisasi model, memastikan bahwa model tersebut tetap tangguh di berbagai tugas dan kumpulan data.

Sebuah alur kerja dua tahap juga telah diperkenalkan untuk mengatasi bias dalam model bahasa yang telah dilatih sebelumnya, dengan fokus pada pengurangan bias dalam konteks internal dan hilir, sekaligus mempertahankan daya ekspresif model. Hal ini dilengkapi dengan kerangka kerja baru yang mengkaji bias dalam model bahasa berbasis transformator yang telah dilatih sebelumnya melalui pemangkasan gerakan, yang menawarkan wawasan baru tentang bias gender dan mengusulkan perbaikan pada metode debiasing yang ada. Selain pendekatan-pendekatan ini, telah diakui bahwa banyak metode debiasing mengabaikan interaksi antara berbagai bias sosial. Untuk mengatasi hal ini, sebuah model debiasing baru telah diusulkan, yang memanfaatkan sinergi antara berbagai bias sosial untuk secara bersamaan memitigasi berbagai bias.

Lebih lanjut, sebuah metode telah diperkenalkan untuk debiasing pembelajaran kontrastif, yang bertujuan untuk meringankan fitur laten yang bias dan mengurangi keberadaannya dalam representasi model. Terakhir, hubungan antara bias ekstrinsik dan intrinsik dalam model NLP masih merupakan area yang relatif belum dieksplorasi. Sebuah kerangka kerja baru telah diusulkan untuk mengukur kedua jenis bias secara bersamaan, menawarkan perspektif yang lebih komprehensif tentang bias dalam model NLP.

BAB 10
MANFAAT DAN RISIKO LLM

10.1 PENDAHULUAN

Kemajuan luar biasa dalam LLM seperti GPT-4 telah mengubah lanskap AI, menunjukkan kemampuan yang belum pernah ada sebelumnya dalam memahami dan menghasilkan teks seperti manusia. Dari iterasi awal mereka, seperti GPT-1, hingga model yang lebih canggih seperti ChatGPT, sistem ini telah menunjukkan kemampuan untuk menangani beragam tugas yang mengesankan, mulai dari menulis esai hingga menghasilkan kode, seringkali dengan instruksi minimal. Namun, terobosan dalam teknologi AI ini menimbulkan kekhawatiran etika yang mendalam. Meskipun LLM menawarkan potensi transformatif di bidang-bidang seperti pendidikan, layanan kesehatan, dan layanan pelanggan, adopsi mereka yang luas juga menimbulkan pertanyaan tentang keandalan, transparansi, dan dampak sosialnya.

Kemampuan utama LLM adalah kemahirannya dalam pembelajaran few-shot, di mana model dapat melakukan tugas-tugas yang belum pernah dilatih secara eksplisit dengan akurasi yang kompetitif. Kemampuan adaptasi ini merupakan langkah menuju AI umum, tujuan jangka panjang dalam komunitas AI. Namun, kemampuan LLM untuk menggeneralisasi ke berbagai domain tanpa penyesuaian juga menimbulkan kekhawatiran etis terkait penyalahgunaan dan konsekuensi yang tidak diinginkan. Tanpa batasan yang ketat, LLM dapat diterapkan di area kritis, seperti nasihat hukum atau dukungan kesehatan mental, di mana keluaran yang salah atau bias dapat menyebabkan kerugian yang signifikan. Fleksibilitas ini, meskipun mengesankan, menuntut kerangka kerja etis yang membahas potensi risiko penggunaan teknologi tersebut dalam lingkungan berisiko tinggi.

Lebih lanjut, skala besar model-model ini, yang dilatih dengan triliunan parameter dan teks dalam jumlah besar, menimbulkan risiko tambahan terkait transparansi, akuntabilitas, dan privasi data. Seiring LLM terintegrasi ke dalam sistem yang lebih sosial, ketidakjelasan cara mereka mengambil keputusan, ditambah dengan ketergantungan mereka pada data pelatihan yang berpotensi bias atau usang, menimbulkan dilema etika yang signifikan. Laju pesat kemajuan AI melampaui perkembangan kerangka kerja regulasi, sehingga menimbulkan kekhawatiran tentang tata kelola teknologi ini yang bertanggung jawab.

Salah satu kemampuan LLM yang paling mengesankan adalah kemampuannya untuk menyelesaikan tugas-tugas baru yang belum pernah dilatihkan tanpa perlu menyesuaikan parameter model dengan tugas baru tersebut. Kemampuan ini dikenal sebagai pembelajaran few-shot. LLM tingkat lanjut adalah pembelajar few-shot, yang berarti mereka dapat menangani tugas-tugas baru dengan kinerja yang kompetitif dibandingkan dengan model yang telah disetel secara halus untuk tugas spesifik tersebut. Kemampuan adaptasi LLM terhadap tugas-tugas baru ini menunjukkan bahwa jenis AI ini adalah yang pertama mencapai terobosan signifikan dalam hal pendekatan AI umum. Meskipun sebagian besar kemajuan ini

disebabkan oleh kapasitas model-model ini yang sangat besar, dengan triliunan parameter, dilatih pada teks dalam jumlah besar, dan diselaraskan dengan instruksi proses berdasarkan umpan balik manusia yang masif, tetap mengesankan bahwa model-model ini terus berkembang pesat. Namun, terlepas dari kemajuan pesat ini, terdapat risiko inheren yang terkait dengan kemajuan pesat teknologi bahasa ini. Kami akan membahasnya di bagian-bagian berikut, yang mengarah pada refleksi tentang aspek etika, penggunaan, dan penyalahgunaan LLM.

10.2 HALUSINASI DALAM LLM

LLM sering menunjukkan kecenderungan untuk menghasilkan halusinasi, sehingga menghasilkan keluaran yang tidak konsisten dengan fakta dunia nyata. Fenomena ini menimbulkan tantangan yang signifikan karena hasil yang andal sangat penting untuk penerapannya dalam berbagai tugas. dalam survei terbaru mereka tentang subjek ini, kami telah mengidentifikasi berbagai jenis halusinasi dalam LLM. Jenis pertama yang kita bahas adalah "halusinasi faktualitas", di mana LLM terkadang menghasilkan keluaran yang tidak konsisten dengan fakta dunia nyata atau berpotensi menyesatkan.

Hal ini dapat disebabkan oleh inkonsistensi faktual, di mana model memberikan keluaran yang didasarkan pada fakta tetapi tidak akurat atau kontradiktif. Ini termasuk kesalahan seperti mengaitkan peristiwa dengan angka yang salah, ketidakakuratan tanggal, dan kesalahan historis lainnya. Halusinasi faktual juga dapat muncul dari fabrikasi faktual, di mana model menciptakan data tentang informasi yang tidak dapat diverifikasi, termasuk mitos urban atau teori konspirasi. Baik inkonsistensi faktual maupun fabrikasi tersebut menyoroti risiko penggunaan LLM sebagai sumber data.

Jenis halusinasi lainnya, yang dikenal sebagai "halusinasi kesetiaan", melibatkan kesalahan dalam logika penyelarasan pertanyaan dan jawaban. Ketidakkonsistenan ini menunjukkan bahwa LLM terkadang gagal memproses perintah dengan benar, sehingga menghasilkan respons yang tidak selaras dengan instruksi pengguna. Kami mengenali tiga jenis halusinasi kesetiaan:

1. Ketidakkonsistenan instruksi, di mana keluaran kurang sesuai dengan maksud pengguna, termasuk kegagalan dalam mengikuti atau salah memahami instruksi.
2. Ketidakkonsistenan konteks, di mana pengguna memberikan fakta dunia nyata dalam konteks perintah, dan keluarannya bertentangan dengan pernyataan ini.
3. Ketidakkonsistenan logika, yang melibatkan kesalahan dalam operasi matematika atau penerapan prinsip-prinsip logika, yang menunjukkan kontradiksi dalam tugas penalaran.

Penyebab halusinasi ini beragam, dengan salah satu faktor signifikan adalah penggunaan sumber data yang tidak konsisten selama pra-pelatihan model. Faktor-faktor yang dapat memperburuk halusinasi termasuk misinformasi yang secara tidak sengaja dimasukkan dalam data pelatihan. Karena LLM menghasilkan keluaran berdasarkan data yang telah diproses selama pelatihan, penyebab-penyebab ini berkaitan dengan kepalsuan imitatif, intinya, jika data mengandung kepalsuan, LLM akan mereproduksinya. Sumber halusinasi lainnya dapat

berupa pengenalan bias sosial dan historis dari sumber data. Ini termasuk "bias duplikasi", di mana fakta yang berulang dalam data dapat menyebabkan LLM beralih dari generalisasi ke hafalan, dengan memprioritaskan mengingat kembali data tersebut. Selain itu, keberadaan bias sosial dan historis dalam data dapat diabadikan melalui asosiasi stereotip.

Salah satu penyebab halusinasi dalam LLM adalah karena batasan pengetahuan. LLM memiliki kemampuan terbatas untuk menangani informasi terkini. Pengetahuan faktual yang ketinggalan zaman menghadirkan tantangan, karena model dasar memiliki batasan waktu dan dapat menjadi usang seiring waktu. Perintah yang melampaui batasan ini dapat memaksa model untuk memberikan jawaban yang dapat dihasilkan dari pemalsuan fakta di luar batasan waktu model. Keterbatasan lain dari model ini berkaitan dengan defisiensi pengetahuan domain, yaitu kurangnya penanganan konsep yang terkait dengan domain tertentu. Masalah ini muncul karena LLM sebagian besar dilatih pada kumpulan data pengetahuan umum, sehingga kemampuannya untuk menjawab pertanyaan terkait domain tertentu terbentur batas domain model.

Penyebab halusinasi lainnya meliputi 'pemanfaatan data yang buruk', di mana korelasi palsu yang tertangkap selama pelatihan menyebabkan ketidakakuratan faktual. Hal ini sering kali merupakan akibat dari 'pintasan pengetahuan', di mana model menekankan statistik kedekatan dan ko-kemunculan dari data pra-pelatihan, yang dapat menyebabkan bias model terhadap asosiasi yang salah, sehingga menyebabkan halusinasi. 'Kegagalan mengingat pengetahuan' juga berkontribusi, terutama dengan ketidakmampuan mengingat pengetahuan berekor panjang, yang mengacu pada kesulitan yang dihadapi LLM dalam menggunakan pengetahuan faktual yang jarang muncul dari data pelatihan.

Penyebab halusinasi signifikan lainnya terkait dengan 'penyalahgunaan data dan pengetahuan parametrik' dalam skenario kompleks, seperti menjawab pertanyaan multi-hop, di mana mesin penalaran harus menggunakan beberapa asosiasi entitas untuk merespons secara akurat. Kompleksitas ini seringkali melampaui kemampuan penalaran LLM yang telah terbentuk selama fase pengambilan, sehingga menyebabkan kesalahan.

Selama fase pelatihan, beberapa faktor memicu halusinasi. ‘Kelemahan arsitektur’ yang terkait dengan representasi searah dalam model bahasa kausal selama pra-pelatihan dapat membatasi pemahaman model, karena konteks krusial mungkin muncul di luar cakupan kiri-ke-kanan langsung. ‘Gangguan perhatian’, di mana model gagal mempertahankan perhatian dalam sekuens yang panjang, juga dapat mengakibatkan halusinasi, karena informasi relevan yang jauh dari token masukan saat ini terabaikan, terutama yang lebih memengaruhi dependensi jangka panjang daripada dependensi jangka pendek.

‘Bias paparan’ selama pelatihan menimbulkan perbedaan antara fase pelatihan dan inferensi dalam model bahasa auto-regresif, yang menyebabkan kesalahan kaskade selama pembuatan token karena ketergantungan pada token yang dihasilkan sebelumnya, alih-alih token kebenaran dasar, yang digunakan model selama pelatihan. Halusinasi selama tahap penyelarasan juga dapat terjadi akibat perintah yang diformulasikan dengan buruk dan 'ketidakselarasan keyakinan', di mana model menghasilkan konten yang tidak selaras dengan

pengetahuan faktual tetapi justru mengikuti pendapat pencatat selama proses RLHF, sebuah fenomena yang dikenal sebagai penjilatan.

Dalam fase inferensi, strategi dekode secara inheren melibatkan pengacakan, yang menimbulkan risiko halusinasi. 'Jebakan kemungkinan', di mana kalimat-kalimat dengan probabilitas tinggi belum tentu berguna, diatasi dengan memperkenalkan keacakan selama dekode untuk menciptakan distribusi probabilitas token yang lebih seragam, mengurangi kemungkinan pengambilan sampel token yang lebih jarang tetapi tidak sesuai konteks.

Hubungan antara pencarian berkas, sebuah strategi dekode yang mengondisikan pengambilan sampel ke token yang menjanjikan, dan halusinasi kurang dieksplorasi tetapi menunjukkan potensi dalam mengurangi risiko halusinasi dengan membatasi ruang sampel ke token yang terhubung kuat dengan token yang mungkin. Penyebab halusinasi lainnya selama inferensi meliputi 'representasi dekode yang tidak sempurna' dan 'perhatian konteks yang tidak memadai', yang memprioritaskan fluiditas teks daripada kesetiaan karena keterbatasan transformator lapisan atas, termasuk 'hambatan softmax', yang membatasi ekspresivitas.

10.3 MITIGASI HALUSINASI DALAM LLM

Mendeteksi informasi faktual palsu dalam LLM merupakan tugas yang menantang. Terdapat dua strategi mitigasi utama: satu melibatkan verifikasi klaim otomatis, dan yang lainnya didasarkan pada estimasi ketidakpastian seputar keluaran model. Setiap strategi menghadirkan tantangan yang signifikan: yang pertama memerlukan akses ke sumber untuk verifikasi, sementara yang kedua bergantung pada interpretasi nilai ketidakpastian yang diberikan model. Mengenai penggunaan fakta eksternal untuk verifikasi, strategi ini biasanya membandingkan fakta yang dihasilkan oleh model dengan fakta yang diambil dari basis pengetahuan.

Namun, pendekatan ini dibatasi oleh keterlambatan pembaruan dalam basis pengetahuan seperti DBPedia. Untuk mengatasi masalah penyelarasan hasil dalam konteks yang sensitif terhadap waktu, perlu menggunakan sumber yang tidak terkurasi seperti sumber daya web, yang secara inheren bias dalam berbagai aspek. Ricardo Baeza-Yates mengidentifikasi, selain sumber bias konvensional seperti bias pengambilan sampel dan algoritmik, bias yang dipicu oleh interaksi pengguna di web, seperti seleksi mandiri dan bias aktivitas. Karena pengguna memengaruhi algoritme rekomendasi web yang dipersonalisasi melalui klik mereka, preferensi mereka mencerminkan bias seleksi diri ini dalam konten yang mereka lihat daring.

Lebih lanjut, berdasarkan interaksi web, yang didominasi oleh segelintir orang, terdapat bias aktivitas yang juga memengaruhi rekomendasi. Baeza-Yates menunjukkan adanya lingkaran setan bias di web. Siklus ini memengaruhi web sebagai sumber data, menjadikannya pilihan yang meragukan untuk memitigasi disinformasi, dan mengalami masalah yang sama dengan LLM: tidak dapat diandalkan, bias, dan didominasi oleh segelintir orang. Strategi mitigasi untuk halusinasi, dalam hal ini, informasi faktual palsu dalam LLM, cenderung bergantung pada perhitungan indikator kepercayaan untuk sumber web yang

digunakan, sebuah pendekatan yang meniru upaya manusia dalam berbagai inisiatif pengecekan fakta.

Mengenai estimasi ketidakpastian sebagai teknik mitigasi disinformasi, strategi ini memperkirakan ketidakpastian konten faktual yang dihasilkan oleh model. Hal ini dapat dilakukan dengan memeriksa status internal LLM, baik melalui probabilitas token maupun entropi. Untuk token, ukuran ketidakpastian adalah probabilitas token minimal. Asumsi yang mendasarinya adalah bahwa probabilitas rendah menunjukkan ketidakpastian model. Dalam hal ini, keluaran LLM dapat digunakan dalam prompt baru, yang menginstruksikan LLM untuk menghasilkan keluaran baru berdasarkan keluaran sebelumnya.

Probabilitas token yang dihasilkan akan mengukur keakraban LLM dengan pengetahuan faktual yang dihasilkan dan dengan demikian membantu kita membuang klaim faktual yang tidak dapat diandalkan. Sayangnya, strategi ini berfungsi di LLM Terbuka, seperti LLama 3, dan tidak di LLM yang hanya dapat diakses melalui panggilan API, seperti yang dari OpenAI, karena LLM ini tidak menyediakan probabilitas token dari keluarannya. Karena keterbatasan ini, beberapa penelitian telah membahas masalah estimasi ketidakpastian dari perspektif perilaku LLM. Salah satu cara untuk melakukannya adalah dengan memformulasikan prompt secara tidak langsung beberapa kali ke LLM, mengevaluasi konsistensi respons sebagai proksi untuk ketidakpastian model.

Strategi ini bergantung pada metode yang digunakan untuk memformulasikan kueri tidak langsung ke model. Dimensi lain dalam menganalisis hasil LLM melibatkan deteksi kesetiaan. Konten yang setia dapat diidentifikasi menggunakan pengklasifikasi kesetiaan. Dalam pengembangan ini, terdapat metode berbasis enumerasi tekstual, yang mencari konsistensi antara dua kalimat yang berurutan. Dalam NLP, enumerasi tekstual adalah tugas yang membantu menentukan apakah dua kalimat yang berurutan terhubung oleh hubungan premis-hipotesis, artinya jika salah satu (hipotesis) merupakan konsekuensi logis dari yang lain (premis). Dalam deteksi kesetiaan, keluaran LLM dapat dipecah menjadi kalimat-kalimat, dan kemudian pengklasifikasi enumerasi tekstual diterapkan untuk menghitung skor kesetiaan.

Alur analisis lain melibatkan penggunaan metrik berbasis tanya jawab, yang penting dalam NLP. Strategi ini melibatkan identifikasi klaim dalam keluaran LLM, kemudian menghasilkan pertanyaan yang selaras dengan klaim tersebut. Pertanyaan-pertanyaan tersebut dirumuskan ulang untuk menghasilkan jawaban dan membandingkannya dengan klaim keluaran asli. Dengan membandingkan skor kecocokan antara klaim dan jawaban, kita dapat menghitung kesetiaan LLM. Keterbatasan strategi ini didasarkan pada aspek-aspek seperti pemilihan klaim, pembuatan pertanyaan, dan tumpang tindih jawaban, yang semuanya memiliki kekuatan dan kelemahan yang memengaruhi keandalan estimasi skor kesetiaan.

Demikian pula, penggunaan metrik berbasis prompt telah menarik perhatian. Lini terbaru ini disebut evaluasi berbasis LLM. Idenya adalah memberikan instruksi yang jelas kepada LLM tentang cara mengevaluasi tugas-tugas tertentu, sehingga model itu sendiri menilai kesetiaannya. Berbagai cara untuk mengevaluasi prompt mencakup penggunaan rantai pemikiran atau membiarkan model menghasilkan evaluasi beserta penjelasannya, yang

semuanya bertujuan untuk menggunakan rantai pemikiran (urutan langkah logis yang diambil untuk menghasilkan hasil) atau penjelasan sebagai bukti untuk menilai kesetiaan LLM.

10.4 LLM YANG MENIRU MANUSIA

Kemampuan LLM yang luas untuk meniru bahasa manusia merupakan kemajuan signifikan dalam AI. LLM secara efektif menangani tugas-tugas seperti penerjemahan mesin, penulisan, parafrase, dan berbagai aktivitas produksi teks. LLM dapat menyesuaikan teks yang dihasilkan berdasarkan profil penulis yang telah ditentukan, menunjukkan keterampilan yang berkaitan dengan peniruan bahasa sehari-hari. Kemajuan ini merupakan pencapaian besar bagi AI.

Namun, sebagaimana diamati, LLM dapat menghasilkan "halusinasi", yang memengaruhi keandalan hasilnya. Selain itu, LLM dapat mereproduksi bias berdasarkan pengetahuan parametrik yang dikodekan dalam model dasarnya, yang seringkali menarik data dari sumber dengan tingkat keandalan yang bervariasi. Versi terbaru ChatGPT menyertakan plugin untuk mengekstrak fakta dari sumber daya web eksternal. Sebagaimana telah dibahas, web memiliki bias, dan banyak sumbernya tidak dapat diandalkan, sehingga meningkatkan risiko mengekstrak informasi faktual palsu dari sumber-sumber tersebut. Hal ini membuat LLM tidak dapat diandalkan dalam hal menghasilkan pengetahuan faktual.

Terlepas dari masalah-masalah ini, potensi penyalahgunaan teknologi-teknologi ini sangat besar. Menurut Ferrara, penerapan LLM yang jahat menghasilkan berbagai jenis kerugian, termasuk kerugian pribadi dan pencurian identitas, kerugian finansial dan ekonomi, serta manipulasi informasi (lihat peta pikiran pada Gambar 10.1). Ancaman-ancaman utamanya meliputi:

1. **Kerugian Pribadi dan Pencurian Identitas**: GenAI berpotensi menghasilkan identitas sintetis. Dengan menggunakan pembangkitan teks, profil pengguna dapat ditulis untuk membuat riwayat palsu. Profil palsu merupakan langkah awal menuju penipuan. Sebagaimana dibahas nanti dalam buku ini, transformator visual dan perluasan lain dari arsitektur ini telah memungkinkan penerapan teknologi generatif pada gambar dan video baru-baru ini. Hal ini mendukung pembuatan profil palsu dengan gambar yang sangat realistis. Kemudahan penggunaan alat-alat ini untuk menghasilkan konten yang dimanipulasi membuka pintu bagi peniruan identitas digital, menggunakan kepribadian yang dikenal melalui GenAI untuk meniru orang. Teknologi ini juga dapat digunakan dalam penipuan telepon melalui peniruan identitas suara.
2. **Kerugian Finansial dan Ekonomi**: Kemampuan luar biasa untuk menghasilkan konten yang tidak dapat diandalkan dalam jumlah besar dapat mempersenjatai pelaku kejahatan, yang dapat membanjiri media sosial dengan informasi yang bias. Hal ini dapat memicu alarm di bidang keuangan, bahkan menyebabkan jatuhnya pasar saham dan memungkinkan manipulasi pasar.
3. **Manipulasi Informasi:** Kemampuan luar biasa untuk menghasilkan konten yang tidak dapat diandalkan dalam jumlah besar dapat mempersenjatai pelaku kejahatan, yang dapat mengoordinasikan kampanye pengaruh terhadap opini publik, membanjiri

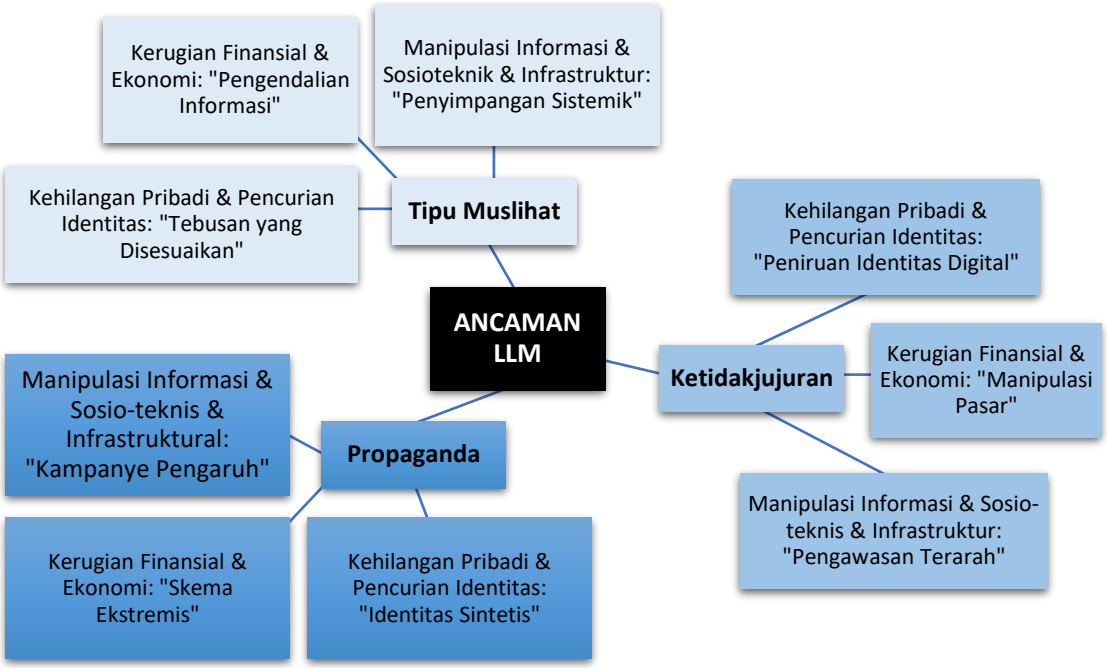
jaringan dengan bot, dan menyebabkan kekacauan informasi. Penggunaan strategi propaganda seperti pengulangan, ajakan kebencian, kekeliruan hitam-putih, dan strategi linguistik persuasif lainnya berada di tangan para pelaku yang dapat menggunakannya untuk memengaruhi opini publik, menciptakan polarisasi dan membangun kontroversi, dengan kendali yang signifikan atas informasi di jejaring sosial.

Salah satu cara untuk mengurangi dampak buruk penyalahgunaan LLM adalah dengan mendeteksi teks yang dihasilkan LLM. Namun, LLM telah berkembang pesat dalam produksi teks sehingga tidak dapat dideteksi oleh manusia. Uji kecerdasan mesin yang terkenal telah lama menjadi uji Turing, yang didasarkan pada prinsip permainan imitasi.

Sebagaimana dinyatakan oleh Alan Turing, jika dua agen, misalnya manusia dan mesin, menghasilkan teks, dan manusia tersebut tidak dapat membedakan mana yang manusia dan mana yang sistem AI, hal ini menunjukkan bahwa AI telah berhasil meniru bahasa manusia sehingga memiliki kemampuan teks generatif yang serupa dengan manusia. Permainan imitasi, yang dikenal sebagai Tes Turing, telah dilampaui oleh Chat-GPT, menunjukkan bahwa manusia tidak dapat membedakan antara teks sintetis dan alami. Kemampuan mimetik ini menyebabkan kesulitan dalam mendeteksi teks yang dihasilkan oleh LLM.

Salah satu pendekatan deteksi melibatkan penggunaan metode algoritmik. Menghasilkan kumpulan data untuk deteksi merupakan hal yang kompleks karena kesulitan yang dihadapi manusia dalam membedakan dan, oleh karena itu, menganotasi serta mengawasi pembuatan kumpulan data untuk tugas ini. Namun, ada beberapa petunjuk yang dapat difokuskan untuk membedakan perbedaan ini. OpenAI telah merilis pengklasifikasi yang dilatih untuk membedakan antara teks yang dihasilkan manusia dan teks sintetis. Pengklasifikasi biner ini dikembangkan berdasarkan teks yang dianotasi dan, meskipun merupakan kemajuan awal yang dibuat oleh pengembang ChatGPT, pengklasifikasi ini memiliki keterbatasan. Khususnya, ia berjuang dengan mengklasifikasikan teks pendek, membuat deteksi teks sintetis pada platform media sosial tidak dapat diandalkan.

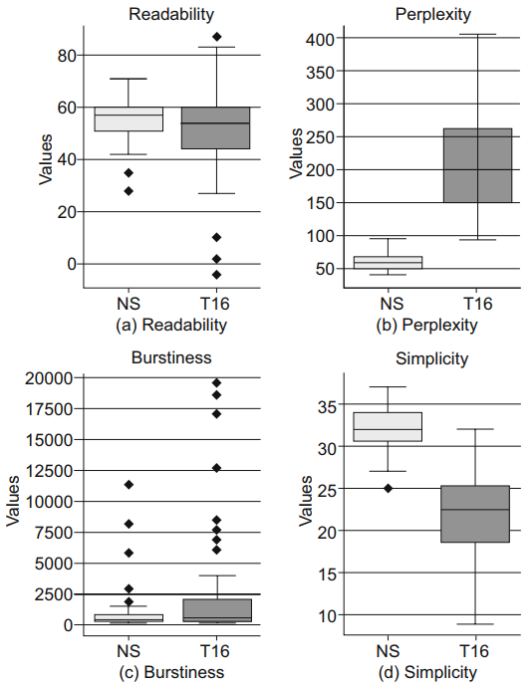
Keterbatasan lain terkait dengan perkiraan yang terlalu tinggi dari teks sintetis; pengklasifikasi cenderung salah mengklasifikasikan teks yang benar-benar dihasilkan oleh manusia sebagai yang dihasilkan AI. Saat ini, hanya tersedia dalam bahasa Inggris. Ia juga tidak dapat membedakan teks sintetis yang berisi informasi yang diverifikasi secara faktual, karena pengklasifikasi membingungkan keluaran yang dapat diandalkan dengan keluaran manusia/buatan. Pengklasifikasi lain yang dirancang untuk tujuan ini adalah GPTZero, yang dilatih untuk mendeteksi teks yang dihasilkan oleh ChatGPT, GPT-4, Bard, LLaMa, dan model AI lainnya. Meskipun hak milik, model ini menyediakan berbagai indikator yang menyoroti petunjuk yang membedakan teks sintetis dari teks yang dihasilkan manusia. Pada Gambar 10.2, kita dapat melihat perbandingan berdasarkan teks yang dihasilkan menggunakan ChatGPT pada model dasar GPT 3.5 versus teks yang dikumpulkan dari percakapan Twitter di situs berita. Dataset Twitter16 relatif lama, menunjukkan rendahnya kehadiran bot, yang memungkinkan atribusi sebagian besar teks ini kepada manusia pada awalnya.



Gambar 10.1: Beberapa ancaman yang ditimbulkan oleh LLM yang meniru manusia yang teridentifikasi di Ferrara.

Gambar tersebut menampilkan tiga metrik yang dihitung oleh GPTZero: keterbacaan, kebingungan, dan ledakan. Keterbacaan mengacu pada seberapa mudah teks dapat dibaca oleh pengguna. Dalam metrik ini, teks sintetis sedikit lebih mudah dibaca daripada teks manusia, kemungkinan karena konstruksi tata bahasa yang lebih akurat yang digunakan oleh LLM, dibandingkan dengan tata bahasa manusia yang seringkali membingungkan dan berbelit-belit, terutama di media sosial. Kebingungan mengukur kompleksitas prediktif suatu teks berdasarkan mekanisme pembangkitan autoregresi. Jelas, kebingungan teks sintetis lebih rendah daripada teks manusia, yang menunjukkan bahwa teks manusia jauh lebih tidak terduga daripada yang dihasilkan oleh LLM.

Perbedaan ini disebabkan oleh strategi pengambilan sampel yang digunakan oleh LLM, yang mengandalkan varian pengambilan sampel acak terkondisi, seperti pencarian berkas. Manusia jauh kurang terprediksi dalam hal ini. Ledakan menilai penggunaan token yang sama secara berulang, efek linguistik yang umumnya dikaitkan dengan penggunaan slogan atau strategi propaganda lainnya. Gambar tersebut menunjukkan bahwa perbedaan karakteristik ini antara kedua jenis teks tersebut sangat kecil. Singkatnya, perbandingan tersebut menunjukkan bahwa perbedaan utama antara teks manusia dan teks sintetis berasal dari cara LLM mengambil sampel kosakata untuk menghasilkan teks mereka, dengan teks yang dihasilkan secara algoritmik jauh lebih dapat diprediksi daripada teks manusia.



Gambar 10.2: Perbandingan antara percakapan alami (T16) dan percakapan sintetis (NS).

Metode lain untuk mendeteksi teks sintetis mencakup penggunaan watermarking algoritmik. Implementasi strategi ini bergantung pada adopsinya oleh perusahaan teknologi besar, seperti OpenAI.

10.5 REFLEKSI TENTANG MANFAAT DAN RISIKO LLM

LLM menawarkan manfaat dan peluang yang signifikan. Kemampuannya yang luar biasa dalam produksi teks dapat menguntungkan jika teknologi ini digunakan untuk tujuan yang baik. Namun, teknologi canggih seperti ChatGPT juga memiliki risiko penyalahgunaan. Di tangan yang salah, teknologi ini dapat menyebabkan kerugian di berbagai bidang. Kerugian ini terutama disebabkan oleh kemampuannya untuk meniru bahasa manusia, sehingga menjadikannya alat yang berpotensi disalahgunakan.

Upaya mitigasi saat ini bergantung pada strategi deteksi, yang saat ini memiliki banyak keterbatasan. Strategi ini memerlukan pengembangan model AI spesifik untuk mendeteksi AI lain, seperti GPTZero. Namun, seiring dengan peningkatan teknologi deteksi, teknologi untuk membuat keluaran tampak lebih mirip manusia pun meningkat. Seiring dengan semakin jelasnya petunjuk deteksi, karakteristik ini dapat disembunyikan secara algoritmik, yang membuat masalah ini semakin sulit. Pada bab selanjutnya, kami akan mengeksplorasi bagaimana perluasan model generatif ke citra semakin memperluas risiko yang terkait dengan GenAI. Kami akan menyimpulkan bahwa memberikan transparansi yang lebih besar kepada model-model ini sangatlah penting dan merekomendasikan beberapa langkah untuk mengatasi risiko inheren dari teknologi ini.

10.6 KESIMPULAN

LLM telah menimbulkan kekhawatiran yang signifikan terkait risiko keselamatan dan keamanan, termasuk pertimbangan etika, halusinasi, dan masalah injeksi cepat. Dalam lingkungan yang mengutamakan keselamatan, seperti layanan kesehatan dan keuangan, penerapan LLM dapat mengakibatkan halusinasi model, yang dapat membahayakan pengguna yang rentan. Selain itu, ancaman keracunan pengambilan data (retrieval poisoning) dalam aplikasi berbasis LLM memungkinkan pelaku kejahatan untuk memanipulasi keluaran, menunjukkan tingkat keberhasilan yang tinggi dalam skenario dunia nyata. Mengintegrasikan LLM ke dalam lingkungan pendidikan juga menghadirkan tantangan etika, seperti masalah privasi data, bias, dan potensi konsekuensi dari penggantian instruktur manusia.

Penelitian terkini secara aktif mengatasi risiko-risiko ini dengan mengkaji implikasi keselamatan dan masalah etika serta mengidentifikasi arah penelitian di masa mendatang untuk mengembangkan aplikasi LLM yang lebih aman dan lebih etis. Salah satu solusi yang diusulkan melibatkan pemanfaatan korpus QA untuk menyelidiki LLM, memanipulasi baik prompt maupun representasi pengetahuan untuk meningkatkan akurasi dalam lingkungan kritis keselamatan.

BAB 11
TRANSFORMATOR VISUAL DAN MULTIMODALITAS

11.1 PENDAHULUAN

AI Generatif telah berkembang pesat dalam menangani format di luar teks. Mungkin format yang menunjukkan hasil paling mengejutkan adalah gambar. Hanya dalam satu dekade lebih, sintesis gambar yang dikondisikan pada suatu kelas, dan kemudian dikondisikan pada suatu prompt, telah berkembang dengan cara yang tak terduga. Dari awal mulanya dengan Jaringan Adversarial Generatif (GAN) hingga Transformator Visual (ViT), fokusnya adalah pada pembuatan gambar beresolusi tinggi. Meskipun sebagian besar model komersial berdasarkan arsitektur ini berfokus pada pembuatan gambar dalam berbagai gaya, pembuatan gambar hiperrealistis mungkin merupakan perkembangan yang paling menggemparkan.

Dalam bab ini, kami akan menjelaskan kemajuan pesat teknologi ini, serta aplikasinya dan ancaman utama yang ditimbulkan oleh potensi penyalahgunaannya.

11.2 JARINGAN ADVERSARIAL GENERATIF

Sintesis gambar telah menjadi salah satu tujuan utama AI dalam dekade terakhir. Salah satu arsitektur pertama yang sukses di bidang ini didasarkan pada GAN, sebuah arsitektur jaringan saraf tiruan yang dirancang untuk memperkirakan parameter model generatif menggunakan proses adversarial di mana dua model dilatih secara bersamaan. Salah satunya adalah model generatif yang menangkap distribusi data pelatihan, yang akan kita sebut sebagai model G, dan model diskriminatif D yang memperkirakan probabilitas bahwa sampel berasal dari data pelatihan.

Kedua model, G dan D, bersaing berdasarkan tujuan mereka, mendefinisikan permainan adversarial yang berfungsi sebagai strategi pelatihan. Sementara D mencoba meminimalkan probabilitas membuat kesalahan, G memaksimalkan probabilitas bahwa D membuat kesalahan. Dalam teori permainan, jenis kompetisi ini dikenal sebagai permainan dua pemain minimax. menunjukkan bahwa dalam ruang fungsi arbitrer G dan D, ada solusi unik sehingga G mengaproksimasi data pelatihan, dan D tidak dapat membedakan antara data sintetis dan data nyata. Artinya, D, dalam masalah klasifikasi biner nyata/palsu (nyata menunjukkan bahwa sampel berasal dari data, palsu menunjukkan bahwa sampel tersebut sintetis), menghasilkan $p =$ untuk sampel apa pun.

GAN awal dilatih menggunakan perceptron multilayer untuk kemudahan, karena memungkinkan penerapan algoritma backpropagation untuk menyimpulkan parameter G dan D. Meskipun melatih kedua model secara bersamaan memerlukan tantangan komputasi, menggunakan strategi bergantian, melatih D selama k langkah sambil menjaga G tetap beku, lalu membekukan D dan melatih G selama satu langkah (istilah "beku" menunjukkan bahwa parameter akan mempertahankan nilainya tanpa perubahan). Misalkan p_g adalah distribusi tak seragam pada sampel yang ditransformasi (sintetis) dan misalkan p_{data} adalah distribusi

data pelatihan. Dapat ditunjukkan bahwa permainan minimax memiliki optimum global ketika $p_g = p_{data}$, dan dalam kasus ini, $D(x)$ = untuk semua x. Dalam praktiknya, untuk menghasilkan contoh-contoh yang ditransformasikan (sintetis) yang memungkinkan G untuk dilatih, dilambangkan dengan variabel z, pengambilan sampel dilakukan dari prior derau $p_g(z)$. Jika G dan D memiliki kapasitas yang cukup, baik untuk mengaproksimasi data maupun untuk mendiskriminasi, maka pg konvergen menjadi pdata.

Dalam praktiknya, jaringan GAN mengalami beberapa masalah, karena konvergensi ke optimum global tidak terjamin karena teorema konvergensi bergantung pada kapasitas G. Oleh karena itu, bergantung pada datanya, sebuah perceptron multilapis mungkin menghasilkan beberapa titik kritis dalam ruang parameter, sehingga sulit untuk mencapai optimum global. Eksperimen awal berdasarkan jaringan GAN digunakan untuk mengaproksimasi digit tulisan tangan (himpunan data AI klasik yang dikenal sebagai MNIST), wajah (Basis Data Wajah Toronto), dan objek atau hewan dalam himpunan data CIFAR 10 (6000 gambar yang sesuai dengan pesawat terbang, mobil, burung, kucing, rusa, anjing, katak, kuda, kapal, dan truk).

Meskipun eksperimen GAN awal ini menunjukkan peningkatan dalam hal metrik (hasil dilaporkan berdasarkan estimasi kemungkinan), persepsi manusia terhadap citra-citra ini masih jauh dari memuaskan. Meskipun citra yang dihasilkan untuk MNIST dan Toronto Face Database (TFD) cukup meyakinkan, hasil untuk CIFAR menampilkan citra dengan warna dan bentuk yang kurang terdefinisi. Baik MNIST maupun (TFD) merupakan kumpulan data dengan keragaman yang lebih rendah dibandingkan CIFAR, baik karena penanganan warna (skala abu-abu) maupun karena beroperasi dalam domain tertentu (digit atau wajah). Beberapa pengujian dalam GANs awal ini mengganti perceptron multilapis dengan jaringan konvolusional, sebuah arsitektur yang dirancang khusus untuk data 2D (seperti citra).

Penggunaan jaringan konvolusional tidak menghasilkan peningkatan signifikan dalam kualitas gambar yang dihasilkan. Salah satu faktor yang berkontribusi terhadap kesulitan dalam menemukan optimum global, dan dengan demikian menyebabkan ketidakstabilan selama pelatihan, adalah penggunaan model skala besar. Meskipun peningkatan kompleksitas parametrik meningkatkan kapasitas model-model ini, hal itu juga mengakibatkan munculnya lebih banyak titik kritis dalam ruang parameter. Dalam pembelajaran mesin, merupakan praktik umum untuk memperkenalkan faktor selama pelatihan yang memberikan penalti atas penggunaan parameter yang berlebihan di setiap langkah pembelajaran. Faktor penalti ini biasanya mencegah model dari overfitting dan umumnya disebut sebagai faktor regularisasi.

Dalam GAN, dari Google DeepMind mengeksplorasi penggunaan regularisasi dalam generator (model G) dengan menerapkan strategi yang dikenal sebagai trik pemotongan. Strategi ini memungkinkan mereka untuk memberikan kontrol yang lebih besar atas trade-off antara fidelitas dan keragaman sampel dengan mengurangi varians input G. Penggunaan regularizer ini memungkinkan model untuk ditingkatkan skalanya, meningkatkan kompleksitas parametrik model generatif, sehingga memungkinkannya untuk beroperasi pada kumpulan data yang lebih besar seperti ImageNet. Model ini, yang dinamai BigGAN, menunjukkan peningkatan signifikan dalam kualitas citra hasil sintesis.

Untuk mengelola trade-off antara fidelitas dan keragaman, BigGAN memperkenalkan trik pemotongan. Strategi ini melibatkan modifikasi prior p(z) yang digunakan untuk pengambilan sampel z. GAN asli mengambil sampel z (data sintetis) dari distribusi Gaussian (khususnya, z diambil dari $N(0,1)$). BigGAN mengambil sampel dari distribusi normal terpotong, di mana nilai-nilai yang berada di luar rentang tertentu, diambil sampelnya kembali hingga berada dalam rentang terkontrol. Penting untuk dicatat bahwa ruang laten bersifat multidimensi, artinya sampel z berkorespondensi dengan vektor stokastik.

Oleh karena itu, secara tegas, trik pemotongan memotong vektor z dengan mengambil sampel ulang nilai-nilai yang besarnya melebihi ambang batas tertentu. Dalam batasnya, ketika ambang batas trik pemotongan mendekati nol, masing-masing sampel mendekati modus distribusi keluaran G. Teknik ini memungkinkan mereka untuk mengelola hiperparameter (ambang batas), yang dapat dikalibrasi post hoc berdasarkan metrik untuk kualitas gambar sintetis, seperti Jarak Awal Frechet (FID). FID menghukum hilangnya keragaman, analog dengan penarikan kembali dalam trade-off presisi/penarikan kembali.

Namun, FID juga menghargai perolehan fidelitas, analog dengan presisi dalam trade-off presisi/penarikan kembali. Untuk memastikan bahwa strategi ini menghasilkan hasil yang baik, G dirancang agar halus sehingga ruang penuh z menghasilkan sampel keluaran yang baik. Untuk mencapai hal ini, Brock dkk. menggunakan varian Regularisasi Ortogonal. Dengan regularizer ini, BigGAN mampu menangani model yang lebih besar yang kompatibel dengan pemotongan dan dapat dikalibrasi berdasarkan FID dengan menyesuaikan ambang batas trik pemotongan. Hasil yang diperoleh BigGAN pada ImageNet sangat luar biasa, tidak hanya karena peningkatan metrik evaluasinya, terutama FID dalam kasus ini, tetapi juga karena resolusi dan kualitas gambar yang disintesis.

Lonjakan kualitas ini kemungkinan merupakan yang pertama yang membawa pembuatan gambar sintetis mendekati hiperrealisme.

11.3 MODEL DIFUSI

Dhariwal dan Nichola dari OpenAI menunjukkan bahwa model difusi mencapai hasil yang lebih baik daripada GAN. Karya penting ini membuka jalur penelitian baru ke dalam model-model ini, yang mendominasi pembuatan citra sintetis selama beberapa tahun. Model difusi adalah kelas model berbasis kemungkinan yang menawarkan properti yang diinginkan seperti cakupan distribusi, tujuan pelatihan stasioner, dan skalabilitas. Perlu dicatat bahwa GAN menunjukkan beberapa masalah ketidakstabilan selama pelatihan, terutama dipicu oleh kurangnya skalabilitas.

Meskipun BigGAN mengatasi tantangan ini, masalah dengan ketidakstabilan dan kesulitan dalam menangani titik kritis selama pelatihan tetap ada, membuat pelatihan jaringan tersebut cukup rumit. Model difusi, di sisi lain, mengatasi tantangan ini melalui pendekatan yang berbeda. Model-model ini menghasilkan sampel dengan menghilangkan noise secara bertahap dari sinyal (data), dan tujuan pelatihannya dinyatakan sebagai batas bawah variasional yang dibobot ulang. Terinspirasi oleh trik pemotongan, Dhariwal dan Nichola menyajikan manfaat menangani trade-off keragaman-fidelitas dalam model difusi.

Secara sederhana, model difusi menggunakan langkah-langkah bertahap untuk mendapatkan sampel yang secara progresif mengandung lebih sedikit derau hingga akhirnya mencapai sampel dari set data pelatihan. Untuk ini, pada setiap langkah waktu, tingkat derau dipertimbangkan, yang melibatkan campuran data nyata dan derau, di mana rasio sinyal terhadap derau ditentukan pada langkah waktu tersebut.

Biasanya, derau tersebut adalah Gaussian. Rantai sampel yang diperoleh dengan mengendalikan rasio sinyal terhadap derau memungkinkan sampel dari data asli yang secara progresif dihilangkan deraunya. Model ini dapat diparameterisasi sebagai fungsi yang memprediksi komponen derau dari suatu sampel, yang berarti fungsi kerugian didefinisikan yang sesuai dengan kesalahan kuadrat rata-rata antara derau sebenarnya dan derau yang diprediksi. Bahasa Indonesia: Untuk mengambil sampel dari prediktor derau, distribusi $p_\theta\ (x_{t-1}|x_t\ )$ dimodelkan sebagai Gaussian yang rata-ratanya dihitung dari prediktor derau dan variansnya tetap (Gaussian bersifat diagonal, sehingga matriks kovariansi bersifat diagonal). Seluruh parameterisasi model difusi awalnya diperkenalkan oleh Ho, yang membuktikan bahwa pendekatan ini memungkinkan sampel berkualitas tinggi ketika jumlah total langkah difusi cukup besar.

Arsitektur yang digunakan oleh Ho. adalah U-Net. U-Net adalah jenis jaringan saraf konvolusional (CNN) yang dirancang khusus untuk tugas segmentasi citra biomedis. Awalnya dikembangkan oleh Olaf Ronnenberger dkk., Philipp Fischer, dan Thomas Brox pada tahun 2015, arsitektur U-Net dicirikan oleh desainnya yang simetris, berbentuk "U", yang memungkinkan transmisi informasi kontekstual yang efisien melalui lapisan-lapisan jaringan. Arsitektur ini terdiri dari dua komponen utama: enkoder yang menangkap konteks gambar dan dekoder yang memfasilitasi pelokalan yang presisi. Koneksi lompat antara enkoder dan dekoder sangat penting, karena keduanya menggabungkan fitur resolusi tinggi dan rendah untuk meningkatkan akurasi segmentasi.

Dari perspektif teknis, enkoder U-Net terdiri dari blok konvolusional yang diikuti oleh operasi pengumpulan maksimum (max-pooling) untuk mengurangi dimensi spasial dan meningkatkan kedalaman fitur yang diekstraksi. Sebaliknya, dekoder memanfaatkan operasi upsampling dan konvolusi untuk merekonstruksi keluaran ke resolusi gambar asli. Koneksi loncatan antara lapisan enkoder dan dekoder yang bersesuaian membantu mempertahankan detail halus pada gambar tersegmentasi, sehingga mengatasi tantangan umum hilangnya informasi dalam jaringan dalam yang terkait dengan masalah gradien menghilang.

Dhariwal dan Nichol memperluas kapabilitas U-Net dengan menggabungkan lapisan atensi pada berbagai resolusi dan meningkatkan jumlah kepala atensi. Dapat disimpulkan bahwa upaya Dhariwal dan Nichol bertujuan untuk menyelaraskan arsitektur U-Net lebih dekat dengan arsitektur Transformer dengan mengintegrasikan beberapa elemen kunci darinya

Aspek penting lain dari karya Dhariwal dan Nichol melibatkan eksplorasi strategi pembangkitan kondisional, yang mereka sebut sebagai panduan pengklasifikasi. GAN menggunakan label kelas selama pembangkitan, mendefinisikan masalah pembangkitan sebagai masalah kondisional kelas di mana diskriminator secara eksplisit dirancang untuk berfungsi sebagai pengklasifikasi.

Dalam model difusi, pengklasifikasi dilatih pada citra bising, dan gradien pengklasifikasi digunakan untuk memandu proses pengambilan sampel difusi menuju kelas arbitrer. Strategi ini menyerupai peran diskriminator dalam GAN, di mana pengklasifikasi bertujuan untuk membedakan antara sampel asli dan palsu. Dalam model difusi, gradien pengklasifikasi sampel yang telah dilatih sebelumnya digunakan untuk memandu proses pengambilan sampel difusi menuju kelas arbitrer. Ini dikenal sebagai proses derau balik bersyarat.

Untuk menerapkan panduan pengklasifikasi pada tugas generatif skala besar, penulis melatih pengklasifikasi di ImageNet. Pengklasifikasi didasarkan pada arsitektur UNet yang sama, dimodifikasi untuk difusi yang stabil, dengan fokus pada lapisan tertentu (lapisan resolusi 8x8), yang menghasilkan keluaran akhir. Pengklasifikasi dilatih pada distribusi denoising yang sama yang digunakan oleh model difusi. Setelah melatih model, pengklasifikasi diintegrasikan ke dalam proses pengambilan sampel difusi.

Salah satu tantangan teknis yang dihadapi adalah penggunaan gradien tanpa penskalaan tidak memungkinkan pengklasifikasi dilatih secara memadai. Namun, dengan menskalakan gradien, pengklasifikasi mencapai hasil yang mendekati optimal. Eksperimen di ImageNet menunjukkan bahwa kombinasi faktor-faktor ini pembangkitan kondisional, panduan kelas, dan penskalaan gradien (dengan faktor 10), memungkinkan para penulis mencapai hasil yang pada saat itu merupakan hasil mutakhir berdasarkan metrik yang dikenal sebagai Skor Awal (IS). Sebaliknya, efek penskalaan kurang signifikan ketika dievaluasi berdasarkan skor FID, yang berarti hasil terbaik diperoleh hanya dengan menggunakan pembangkitan kondisional dan panduan kelas.

Hasil ini terkait dengan fakta bahwa FID memerlukan pengoptimalan keragaman dan fidelitas, sedangkan IS hanya mengukur fidelitas. Eksperimen juga menunjukkan bahwa panduan kelas lebih unggul daripada BigGAN dalam mengatasi tradeoff keragaman-fidelitas. Dari perspektif persepsi manusia, sampel yang dihasilkan oleh BigGAN menunjukkan distorsi pada penanda realisme, seperti geometri wajah. Sebaliknya, model difusi menunjukkan konsistensi yang lebih besar terhadap batasan fisik, baik dalam penanganan bentuk maupun warna, sehingga mendekati sintesis citra hiperrealistis dengan lebih baik.

Salah satu keterbatasan model ini adalah bahwa panduan kelas, elemen kunci dalam menghasilkan citra hiperrealistis, bergantung pada keberadaan data berlabel, sehingga kurang cocok untuk skenario tanpa pengawasan. Solusi potensial yang diantisipasi oleh penulis adalah menghasilkan label sintetis berdasarkan sampel pengelompokan.

11.4 PEMBANGKITAN GAMBAR YANG DIKONDISIKAN PADA TEKS

Perkembangan alami dalam model gambar sintetis adalah pengembangan pembangkit gambar yang dikondisikan pada prompt. Dalam konteks ini. di OpenAI mengatasi tantangan ini dengan mengembangkan model CLIP (Pra-pelatihan Bahasa-Gambar Kontrastif). Kunci CLIP terletak pada bekerja dengan teks dan gambar yang selaras, yaitu teks deskriptif dari suatu gambar (keterangan gambar), melatih model yang memprediksi pasangan yang tepat dari sekumpulan (gambar, teks). Model pasangan ini telah dilatih sebelumnya untuk selanjutnya digunakan dalam model visual yang dapat ditransfer.

Untuk melatih pencocok pasangan, beberapa teks dan gambar digunakan, menyelaraskan gambar dengan keterangannya berdasarkan maksimalisasi produk enkode teks dan gambar. Setelah pasangan teks-gambar dilatih, kumpulan data dibuat dari keterangan tersebut dengan mengganti kata kunci dengan opsi alternatif (meniru apa yang dilakukan dalam model bahasa bertopeng). Misalnya, jika keterangan gambar adalah "Foto seekor anjing", strategi model bahasa bertopeng menghasilkan contoh seperti "Foto sebuah objek", di mana objek dapat diwujudkan dengan berbagai cara (misalnya, pesawat, mobil, anjing, ... burung).

Strategi substitusi masking ini menghasilkan N sampel dari keterangan gambar, yang hanya satu yang benar (anjing). Dengan menggunakan pasangan teks-gambar, diharapkan keterangan yang benar dihasilkan, dan keterangan alternatif dibuang berdasarkan encoder gambar. Model ini beroperasi dalam mode "prediksi zero-shot", karena dapat dilihat bahwa kita telah mentransfer dari encoder gambar ke ruang teks menggunakan data tanpa pengawasan atau jenis strategi transfer eksplisit lainnya. Jenis model ini dikenal sebagai model dengan pengawasan lemah, karena dengan menggunakan trik keterangan gambar, dimungkinkan untuk memiliki data berpasangan dalam jumlah besar menggunakan sumber publik seperti Wikipedia. Pembuat CLIP bekerja dengan kumpulan data yang terdiri dari hampir 400 juta pasangan keterangan gambar.

Hasil eksperimen berdasarkan model ini menunjukkan bahwa, serupa dengan GPT, CLIP mempelajari berbagai tugas, kali ini secara bimodal, dan oleh karena itu berguna untuk OCR, geolokasi, pengenalan tindakan dan postur, di antara tugas-tugas umum lainnya dalam bidang visi komputer. Evaluasi transferabilitas CLIP di hampir 30 set data yang ada dalam bidang visi komputer menggambarkan fleksibilitas model ini dalam berbagai tugas, bahkan menunjukkan keunggulannya dibandingkan model yang dilatih secara khusus pada masing-masing set data tersebut.

Tidak diragukan lagi, CLIP merupakan kemajuan fundamental dalam model generatif bimodal, dalam hal ini, dalam modalitas teks-gambar. Dalam hal arsitektur, CLIP bereksperimen dengan berbagai encoder untuk gambar dan teks. Disimpulkan bahwa transformer adalah arsitektur yang paling cocok untuk mengodekan teks, sedangkan arsitektur ResNet digunakan sebagai encoder gambar. ResNet (Residual Networks) adalah arsitektur jaringan saraf tiruan dalam yang diperkenalkan oleh He dkk. pada tahun 2016 yang memfasilitasi pelatihan jaringan yang jauh lebih dalam untuk gambar dengan menggunakan koneksi lompat.

Koneksi ini memungkinkan sinyal masukan mengalir langsung ke lapisan yang lebih dalam, sehingga mengurangi masalah gradien menghilang yang sering terjadi pada jaringan yang sangat dalam. Komponen utama ResNet adalah blok residual, yang mengintegrasikan masukan blok dengan keluaran serangkaian transformasi menggunakan koneksi lewati yang menambahkan masukan asli ke keluaran lapisan konvolusional perantara. Arsitektur ini terbukti efektif tidak hanya dalam meningkatkan akurasi dalam tugas klasifikasi dan deteksi gambar, tetapi juga dalam mempercepat proses pelatihan CLIP.

Arsitektur kedua yang digunakan dalam CLIP untuk mengodekan gambar adalah Visual Transformer, sebuah arsitektur yang diusulkan oleh Dosovitskiy. Arsitektur Visual Transformer (ViT) menandai inovasi signifikan dalam bidang visi komputer dengan mengadaptasi arsitektur transformator untuk analisis gambar. ViT beroperasi dengan membagi gambar menjadi beberapa patch, yang kemudian diratakan dan diubah menjadi urutan data, mirip dengan cara kata-kata ditangani dalam pemrosesan teks.

Patch ini diproses melalui beberapa lapisan transformator yang menggunakan mekanisme atensi untuk menangkap ketergantungan global di antara mereka. Berbeda dengan arsitektur konvolusional seperti ResNet yang berfokus pada area lokal, transformer memiliki kemampuan untuk memperhatikan bagian mana pun dari gambar berkat kemampuannya mengonfigurasi strategi atensi global. Hal ini memberikan pendekatan yang lebih fleksibel dan berpotensi lebih canggih untuk tugas-tugas seperti klasifikasi gambar dan deteksi objek.

Ciri khas ViT adalah ketergantungannya pada mekanisme atensi, yang memungkinkan model untuk menimbang berbagai bagian gambar berdasarkan relevansinya dengan tugas yang sedang dikerjakan. Hal ini dicapai dengan menghitung skor atensi yang memandu fokus model ke interaksi paling signifikan antar-patch. Karena transformer tidak secara induktif memasukkan bias spasial seperti CNN, transformer membutuhkan data dan daya komputasi yang besar untuk pelatihan yang efektif. Mengingat CLIP dilatih pada dataset yang sangat besar, yaitu 400 juta pasangan gambar-teks, penggunaan ViT dalam pelatihan CLIP sangatlah tepat.

Para pembuat CLIP melaporkan tidak ada perbedaan signifikan dalam hal waktu yang dibutuhkan untuk melatih CLIP. Versi berbasis ResNet membutuhkan waktu sekitar 18 hari pada 592 GPU V100, sementara versi berbasis ViT membutuhkan waktu 12 hari pada 256 GPU V100. Hal ini menunjukkan bahwa ViT menawarkan keunggulan dalam efisiensi pelatihan dibandingkan ResNet. Hasil yang dilaporkan menunjukkan bahwa ViT mencapai kinerja yang superior, sehingga implementasi akhir CLIP didasarkan pada arsitektur ViT.

11.5 MODEL DIFUSI DENGAN TRANSFORMATOR

Keberhasilan arsitektur ViT selama pengembangan CLIP mendorong kemajuan teknologi signifikan berikutnya dalam model generatif citra: pengembangan model difusi berdasarkan arsitektur transformator. Kemajuan ini, yang dipelopori oleh Peebles dan Xie (dengan Peebles melakukan pekerjaan ini dalam grup FAIR di META AI), menggantikan arsitektur U-Net yang sebelumnya digunakan dalam model difusi stabil dengan arsitektur ViT yang sukses, yang telah menunjukkan hasil yang menjanjikan dalam CLIP. Pada dasarnya, gagasan difusi stabil diimplementasikan pada ViT berdasarkan patch laten. Perkembangan ini menunjukkan bahwa bias induktif U-Net tidak krusial bagi kinerja model difusi dan dapat digantikan dengan transformator. Model baru ini, yang mengintegrasikan model difusi dengan transformator, dinamai Transformator Difusi (DiT).

Modifikasi tambahan yang diperkenalkan oleh DiT adalah bahwa ia tidak beroperasi secara langsung pada patch citra, seperti yang dilakukan ViT asli. Sebaliknya, ia mengadopsi strategi yang diperkenalkan oleh Rombach, yang beroperasi pada patch dari ruang laten.

Model-model ini berfungsi dengan mereduksi dimensionalitas citra ke ruang laten berdimensi lebih rendah, tempat proses difusi dilakukan secara lebih efisien. Model difusi belajar untuk secara bertahap menghasilkan sampel baru dari distribusi derau, menyempurnakan sampel-sampel ini secara iteratif di ruang laten sebelum memetakannya kembali ke ruang resolusi tinggi, sebuah strategi yang didasarkan pada Variational Autoencoder (VAE).

Secara teknis, proses ini dimulai dengan mengodekan citra resolusi tinggi ke dalam representasi laten terkompresi menggunakan autoencoder. Kunci dari pendekatan ini adalah, di ruang laten, sebuah model difusi diterapkan, dilatih untuk memodelkan distribusi representasi laten. Model difusi ini secara bertahap membalikkan proses derau yang ditambahkan ke representasi laten, memungkinkan pembuatan citra baru dengan mendekode sampel laten yang telah disempurnakan. Teknik ini tidak hanya mengurangi biaya komputasi secara signifikan dibandingkan dengan model difusi yang beroperasi langsung di ruang citra resolusi tinggi, tetapi juga mempertahankan atau bahkan meningkatkan kualitas citra yang dihasilkan.

Kombinasi kedua strategi, transformator yang diterapkan pada patch dan proses difusi dari ruang laten, merupakan dua faktor kunci yang membuat DiT berfungsi. Meskipun penggunaan transformator memungkinkan hasil resolusi tinggi, penerapan proses difusi di ruang laten memungkinkan DiT untuk diskalakan dalam hal jumlah data yang dapat ditanganinya.

Kemajuan terbaru dalam arsitektur ini melibatkan penggabungan modifikasi akhir pada model DiT: aliran yang direktifikasi. Aliran yang direktifikasi adalah formulasi generatif yang menghubungkan data dan derau dalam garis lurus. Ide di balik strategi ini adalah untuk meningkatkan teknik pengambilan sampel derau yang sudah ada yang digunakan selama proses difusi, sehingga menghasilkan skala yang relevan secara perseptual. Selain itu, model ini, yang disebut sebagai transformator aliran yang direktifikasi, memodifikasi arsitektur DiT untuk sintesis teks-ke-gambar dengan menggunakan bobot terpisah untuk teks dan gambar, sehingga memungkinkan aliran informasi dua arah antara gambar dan token. menunjukkan bahwa modifikasi ini meningkatkan pemahaman teks dalam gambar dan sintesis tipografi, yang semuanya menyiratkan peningkatan persepsi manusia terhadap gambar yang dihasilkan.

Transformator aliran yang disearahkan menggunakan enkoder berbasis CLIP untuk menangani teks. Arsitektur ini juga menambahkan enkoder T5 (enkoder-dekoder transformator teks-ke-teks), yang membentuk representasi gabungan dari data teks masukan. Di sisi lain, enkode yang diperoleh dari CLIP dimasukkan ke dalam MLP dan digabungkan dengan enkode yang mewakili langkah waktu.

Hal ini memungkinkan arsitektur, selain memproses teks sebagai rangkaian simbol, untuk melacak urutan simbol-simbol ini yang disajikan kepada transformator. Arsitektur ini juga menggabungkan laten berderau untuk patch gambar, yang melewati lapisan linier dan ditambahkan dengan enkode posisional untuk merekam posisi patch di dalam gambar. Singkatnya, arsitektur ini menghasilkan tiga masukan untuk blok transformator: masukan dari teks (diproses oleh CLIP dan T5), langkah waktu yang dikombinasikan dengan enkode teks yang

diekstrak dari CLIP, dan enkode patch gambar yang dikondisikan pada laten berderau, yang mencakup penyematan posisional. Ketiga masukan ini dimasukkan ke dalam transformator.

Ketika penyisipan patch gambar dan penyisipan keterangan memasuki blok transformator, penyisipan keterangan yang dikombinasikan dengan pengodean langkah waktu digunakan untuk mengumpankan koneksi residual transformator. Dengan demikian, koneksi loncat ini masuk pada tingkat pemrosesan yang berbeda yang dilakukan oleh setiap blok transformator. Mengenai blok transformator, keduanya secara independen memproses penyisipan patch gambar dan penyisipan keterangan, masing-masing memasuki lapisan linier. Kedua masukan ini digabungkan untuk menghasilkan masukan QKV yang digunakan oleh lapisan perhatian mandiri transformator. Kombinasi ini dilakukan melalui perkalian per elemen. Dari modul perhatian, dua salinan pengodean keluaran diekstraksi, yang dimasukkan ke dalam blok transformator berikutnya, mengulangi proses tersebut.

Arsitektur ini mencapai hasil mutakhir pada ImageNet menggunakan metrik FID. Salah satu wawasan yang diperoleh dari evaluasi arsitektur ini adalah peningkatan kemampuannya dalam menangani keterangan. Peningkatan ini tidak hanya meningkatkan resolusi gambar yang dihasilkan tetapi juga menyediakan keterangan dengan kualitas yang lebih tinggi, baik dalam pengambilan maupun sintesis. Semua aspek ini meningkatkan kemampuan pemahaman bahasa model, memungkinkannya untuk menangkap konten prompt secara lebih efektif saat menghasilkan gambar baru. Hasil yang disajikan melampaui hasil yang dicapai oleh model lain, bahkan ketika mempertimbangkan model komersial seperti DALL-E OpenAI, yang didasarkan pada difusi.

11.6 INTEGRASI MULTIMODA DALAM MODEL PERCAKAPAN

Kemajuan yang ditunjukkan oleh integrasi transformer dengan model difusi stabil begitu luar biasa sehingga memacu pengembangan model yang menggabungkan pemahaman bahasa bimoda. Di antara model komersial, mungkin yang paling menonjol adalah model yang dikembangkan oleh OpenAI, dengan model unggulan mereka adalah GPT-4. GPT-4 memiliki kemampuan untuk memproses teks dan gambar secara bersamaan, membuka berbagai macam aplikasi. GPT-4 awalnya diluncurkan untuk memproses teks dan gambar sebagai masukan dan menghasilkan teks sebagai keluaran.

Kekuatan utamanya terletak pada pembangkitan bahasa alami dalam skenario yang lebih kompleks, dengan prompt yang mengintegrasikan teks dan gambar. Hasil yang mengejutkan adalah bahwa GPT-4 tidak hanya menunjukkan kemampuan untuk menangani data bimoda tetapi juga menunjukkan peningkatan yang signifikan dalam pemahaman bahasa. Hal ini disebabkan oleh kombinasi faktor-faktor yang menguntungkan GPT-4, yaitu, menggabungkan lebih banyak data dengan menggunakan data bimoda dan juga meningkatkan ukuran model.

Mengenai ukuran model, model-model ini telah berkembang pesat. Sementara GPT-2 pada tahun 2019 memiliki 1,5 miliar parameter, GPT-3 pada tahun 2020 meningkat menjadi 175 miliar parameter. GPT-3 dan versi 2022-nya, GPT-3.5, membentuk dasar ChatGPT, yang dirilis pada November 2022. Dari versi-versi ini, yang beroperasi sekitar urutan 175 miliar

parameter, GPT-4, yang dirilis pada Maret 2023, tumbuh menjadi hampir 1000 miliar parameter. Versi GPT-4 yang efisien, GPT-4 Turbo, dirilis pada November 2023.

Diketahui bahwa GPT-4 adalah model yang lebih besar dalam hal kapasitas parameter dan persyaratan komputasi, yang memungkinkannya untuk memahami dan menghasilkan teks dengan tingkat kompleksitas dan fidelitas yang lebih tinggi. Di sisi lain, GPT-4 Turbo dioptimalkan untuk menawarkan respons yang lebih cepat dan penggunaan sumber daya komputasi yang lebih efisien, menjadikannya versi yang lebih gesit dan mudah diakses secara ekonomis, tetapi berpotensi dengan lebih sedikit kemampuan dalam tugas-tugas yang melibatkan pembuatan teks dan pemahaman kontekstual dibandingkan dengan GPT-4. Versi "Turbo" ini dirancang untuk aplikasi yang membutuhkan latensi rendah dan kinerja cepat, ideal untuk lingkungan yang mengutamakan kecepatan respons. Oleh karena itu, versi Turbo diyakini memiliki parameter yang lebih sedikit dibandingkan versi GPT-4 yang dirilis pada Maret 2023.

Pada tahun 2024, sistem percakapan kembali menghadirkan banyak kejutan. Integrasi multimoda, yang telah dieksplorasi dalam GPT-4 berbasis arsitektur ViT, diperluas untuk mencakup modalitas lain, khususnya audio. Penggunaan model pemahaman ucapan merupakan kemajuan signifikan dalam sistem jenis ini. Hal ini telah dieksplorasi dalam berbagai sistem perintah suara, terutama pada asisten Alexa (Amazon) dan Siri (Apple). Inti dari sistem ini adalah model Pengenalan Ucapan Otomatis (ASR), yang mengubah sinyal audio suara manusia menjadi teks.

Sistem ASR modern biasanya menggunakan jaringan saraf tiruan untuk menangkap dependensi temporal dan fitur spasial audio. Setelah audio diubah menjadi teks, LLM digunakan untuk tugas pemahaman bahasa. Penggabungan kedua teknologi, ucapan-ke-teks dan LLM, merupakan kombinasi yang bermanfaat. Faktanya, para pengembang LLM telah mengintegrasikan teknologi ini untuk memfasilitasi interaksi dengan manusia. Misalnya, OpenAI telah mengembangkan model ucapan-ke-teks yang sangat sukses bernama Whisper. Model Whisper adalah sistem Pengenalan Ucapan Otomatis (ASR) yang berbasis pada arsitektur transformator.

Model ini terkenal karena kemampuannya untuk mentranskripsikan teks dari audio dalam berbagai bahasa dan dialek dengan akurasi tinggi. Model ini menggunakan transformator perhatian penuh, yang memproses urutan audio untuk memprediksi transkripsi teks yang sesuai. Kunci efektivitasnya terletak pada pelatihannya, yang dilakukan menggunakan kumpulan data yang sangat besar dan beragam yang mencakup berbagai bahasa, aksen, dan kondisi akustik. Pendekatan ini memungkinkan model untuk menangani variasi linguistik dan akustik secara efektif, menjadikannya alat yang tangguh untuk aplikasi transkripsi dalam berbagai konteks dan lingkungan.

Whisper menerapkan pendekatan pengodean berbasis token yang mengubah masukan audio menjadi representasi laten, yang kemudian diproses oleh blok transformator untuk memprediksi urutan teks. Selain itu, model ini memanfaatkan mekanisme atensi yang memungkinkannya berfokus pada bagian tertentu dari masukan audio, sehingga meningkatkan akurasi transkripsi. Implementasi praktis Whisper juga mencakup fitur-fitur

seperti deteksi bahasa otomatis dan kemampuan untuk menangani audio berkualitas rendah. Fleksibilitas ini mendorong pengembangan model terbaru OpenAI, GPT-4o ("o" berarti "omni"), sebuah model menyeluruh yang menggabungkan pemahaman gambar, audio, dan bahasa, yang dibangun di atas GPT-4. GPT-4o diluncurkan pada Mei 2024 dan tersedia untuk versi desktop, akses API, dan akses seluler.

META tidak ketinggalan. Melalui LLM-nya, LLama telah mengambil posisi penting di ranah LLM. Keunggulan utama META dibandingkan OpenAI adalah operasinya di platform media sosial. Mengingat META mengelola platform seperti Facebook, Instagram, dan WhatsApp, integrasi model Llama ke dalam platform ini merupakan perkembangan yang wajar. Meta memang telah mengaktifkan LLM-nya di akun WhatsApp bernama Meta AI. Meta AI beroperasi pada LLM Llama 3 dan dapat diaktifkan di aplikasi WhatsApp kami sebagai kontak tambahan dalam daftar kontak kami. Aktivasi Meta AI di WhatsApp telah tersedia sejak Juli 2024 di semua ponsel kami.

Perusahaan teknologi besar lainnya yang telah berupaya keras untuk mengimbangi perkembangan di arena LLM adalah Google. Google telah menjadi pemain kunci dalam perkembangan teknologi informasi selama dua dekade terakhir. Sejak awal berdirinya, produk andalannya adalah mesin pencari. Mesin pencari Google memimpin era Web 1.0, menjadi mesin pencari yang paling kuat dan paling banyak diadopsi di Barat. Dalam lingkungan yang sangat dinamis, kemunculan jejaring sosial mendorong peluang bisnis baru di bidang teknologi, dan pemain lain memasuki ranah teknologi pada dekade kedua abad ke-21.

Kebangkitan Facebook dan kemudian Instagram, yang kini berada di bawah META, telah memperumit lanskap teknologi. Persaingan ketat di antara para raksasa ini telah memacu kemajuan lebih lanjut dan memicu investasi signifikan dalam pengembangan teknologi. Dominasi AI tidak diragukan lagi telah menjadi landasan persaingan ini selama lima tahun terakhir. Dalam konteks ini, Google telah mengembangkan lingkungan Google AI-nya, yang menyediakan perangkat AI bagi para pengembang. Relevansi Google dalam NLP sangat mendasar dalam sistem penerjemahan mesin, dan Google Translate kemungkinan merupakan sistem NLP pertama yang mencapai adopsi luas secara global. Namun, masuknya Google ke dalam persaingan LLM agak tertunda.

Mengingat GPT pertama muncul tak lama setelah diperkenalkannya arsitektur Transformer, yang dapat ditelusuri kembali ke tahun 2018, peluncuran Bard memang terlambat. Google meluncurkan model AI-nya yang dikenal sebagai Bard pada Maret 2023. Bard didasarkan pada arsitektur Language Model for Dialogue Applications (LaMDA), yang dirancang khusus untuk meningkatkan pembuatan teks dalam konteks percakapan. Model ini memanfaatkan teknik pembelajaran mendalam untuk memahami dan menghasilkan bahasa manusia secara lebih alami dan kontekstual. Papan peringkat LLM, berdasarkan evaluasi di berbagai tolok ukur seperti MMLU (Massive Multitask Language Understanding), secara konsisten menunjukkan bahwa Bard tertinggal jauh di belakang model-model OpenAI.

Google telah melakukan upaya signifikan untuk kembali meraih posisi di area ini, berinvestasi dalam model percakapan barunya, Gemini. Diluncurkan pada 6 Desember 2023, model ini merupakan kemajuan besar dalam bidang model bahasa AI. Dikembangkan sebagai

respons terhadap kemajuan yang dicapai oleh model generatif seperti GPT, Gemini menonjol karena arsitektur atensi dua arahnya dan kemampuan pelatihannya yang efisien di berbagai tugas linguistik dan pemrosesan data multimodal. Model ini dibangun di atas teknik pembelajaran mendalam yang canggih, termasuk pengoptimalan arsitektur transformator dan teknik pembelajaran terfederasi.

Mengingat relevansinya dengan pelatihan Gemini, penting untuk menjelaskan pembelajaran terfederasi secara singkat. Pembelajaran terfederasi adalah paradigma pembelajaran mesin yang memungkinkan pelatihan model AI terdistribusi dengan tetap menjaga privasi data. Dalam pendekatan ini, beberapa server berpartisipasi dalam pelatihan model global tanpa membagikan data lokal mereka. Setiap node melatih salinan model menggunakan set datanya sendiri, lalu hanya mengirimkan pembaruan parameter model ke server pusat. Server ini mengumpulkan pembaruan untuk menyempurnakan model global, yang selanjutnya didistribusikan ke semua node untuk iterasi selanjutnya. Metode ini tidak hanya membantu melindungi privasi tetapi juga mengurangi kebutuhan untuk mengirimkan data dalam jumlah besar.

Pemain kunci lain dalam AI percakapan adalah Microsoft. Minatnya yang kuat, ditambah dengan kolaborasi erat dengan OpenAI, telah memungkinkannya untuk mendapatkan tempat dalam persaingan LLM. Lingkungan Azure-nya tidak diragukan lagi merupakan produk unggulannya bagi para pengembang solusi, yang secara menonjol menampilkan akses tertanam ke GPT-4. Pada tahun 2023, aliansi ini telah mulai membuahkan hasil dengan integrasi LLM ke Azure, dan selanjutnya ke beberapa produk unggulan Microsoft, seperti mesin pencari Bing. Integrasi LLM ke dalam mesin pencari Bing patut mendapat perhatian khusus. Bing telah mengintegrasikan teknologi AI canggih, khususnya GPT-4 OpenAI, untuk meningkatkan kemampuan pencarian dan responsnya. Integrasi ini dicapai dengan menghubungkan mesin pencari dengan API AI, yang memungkinkan Bing untuk memproses kueri bahasa alami dan menghasilkan respons yang tidak hanya relevan tetapi juga akurat secara kontekstual. Model AI yang dilatih pada kumpulan data yang ekstensif dapat memahami dan menghasilkan bahasa alami, memfasilitasi interaksi yang lebih intuitif dan efisien dengan pengguna.

Selain itu, integrasi kemampuan pembelajaran mesin berkelanjutan memungkinkan Bing untuk beradaptasi dan meningkatkan responsnya berdasarkan interaksi pengguna dan perubahan informasi yang tersedia di web. Integrasi kapabilitas mesin pencari Bing dengan pemahaman bahasa GPT-4 tidak diragukan lagi merupakan aspek paling menonjol dari kemitraan antara Microsoft dan OpenAI ini.

Microsoft juga telah menjajaki pengembangan LLM. Model unggulannya, yang dikenal sebagai model Phi, khususnya terkenal karena jumlah parameternya yang lebih kecil. Microsoft telah mengembangkan model yang lebih kecil daripada OpenAI, terutama karena tokenisasi dan pembagian parameter yang efisien, yang mengurangi redundansi sekaligus mempertahankan atau bahkan meningkatkan kinerja model dalam tugas NLP. Lebih lanjut, Microsoft menggabungkan metode kuantisasi dan pemangkasan selama dan setelah pelatihan, menghasilkan model yang lebih ringan dan lebih cepat. Pendekatan ini

memungkinkan model Phi untuk menangani data bervolume besar dan operasi inferensi kompleks secara lebih efisien, yang krusial untuk aplikasi waktu nyata dan perangkat dengan sumber daya terbatas, seperti perangkat seluler.

11.7 PEMAIN BARU MEMASUKI BIDANG AI PERCAKAPAN

Perkembangan menarik dalam persaingan LLM adalah munculnya pemain yang merupakan pihak luar bagi perusahaan-perusahaan besar. Bahkan, OpenAI pun dapat dianggap sebagai pihak luar, dengan asal-usul yang awalnya lebih dekat dengan inisiatif akademis daripada perusahaan besar. Dalam konteks ini, perusahaan seperti Anthropic dan Perplexity patut dikaji. Anthropic adalah perusahaan AI yang didirikan pada tahun 2021 oleh mantan anggota tim OpenAI, termasuk Dario Amodei, yang merupakan Wakil Presiden Riset di OpenAI.

Perusahaan ini didirikan dengan fokus pada pengembangan teknologi AI yang lebih aman dan etis, yang bertujuan untuk mengatasi isu-isu mendasar terkait penyelarasan dan tata kelola model AI skala besar. Produk-produk awal Anthropic meliputi Claude, sebuah model bahasa skala besar yang dirancang agar mudah diinterpretasi dan lebih kecil kemungkinannya menghasilkan konten yang berbahaya atau menyesatkan. Model ini dikembangkan sebagai alternatif dari model yang lebih dikenal seperti GPT-3, dengan peningkatan spesifik dalam hal ketahanan dan keamanan melalui teknik-teknik seperti interpretasi kausal dan koreksi konstitutif dalam pelatihan model.

Baru-baru ini, Anthropic merilis model Claude 3.5 Sonnet, sebuah LLM dengan kemampuan pemahaman bahasa yang substansial, yang mendekati hasil yang dapat dicapai menggunakan GPT-4. Model ini merupakan bagian dari iterasi berkelanjutan untuk meningkatkan keamanan dan keselarasan model bahasa skala besar, dengan menggabungkan inovasi dalam pelatihan dan desain model. Berdasarkan kemajuan Claude sebelumnya, model ini mengintegrasikan teknik-teknik spesifik untuk memitigasi risiko seperti pembuatan informasi palsu dan penguatan algoritma penyaringan untuk menghindari konten yang tidak diinginkan. Melalui pendekatan pelatihan iteratif dan terawasi, Claude 3.5 Sonnet meningkatkan aspek-aspek penting seperti pemahaman konteks dan pembuatan respons yang lebih koheren dan sesuai konteks.

Pengembangannya menandai tonggak penting dalam upaya menciptakan sistem AI yang tidak hanya tangguh dalam kemampuan linguistik tetapi juga lebih aman dan lebih selaras dengan harapan etika dan sosial. Perplexity.ai, yang lebih baru, diluncurkan pada tahun 2023 sebagai perusahaan rintisan di bidang AI, yang berfokus pada pengembangan dan komersialisasi model bahasa tingkat lanjut. Perplexity.ai lebih mendekati apa yang kita kenal sebagai mesin pencari yang terintegrasi dengan LLM, mirip dengan integrasi Bing dengan GPT oleh Microsoft. Perplexity lebih menekankan penanganan URL, lebih menyerupai apa yang diamati di mesin pencari tradisional. Perplexity secara eksplisit menyatakan sumber-sumber ini dalam responsnya, sehingga berkontribusi pada transparansi dan keterlacakan data, sesuatu yang umumnya sulit dicapai di LLM lain.

11.8 HIPERREALISME DALAM GERAKAN

Kemajuan terbaru dalam AI generatif telah memungkinkan terciptanya gambar bergerak hiperrealistis. Meskipun spesifikasi teknis di balik perkembangan impresif ini sebagian besar masih belum diketahui, terdapat indikasi bahwa model seperti Sora untuk pembuatan video dari teks juga dapat mengandalkan arsitektur ViS atau DiS. Namun, alih-alih beroperasi pada patch gambar, model ini bekerja pada rangkaian patch, yang urutannya menunjukkan saling ketergantungan temporal. Hal ini dibahas dalam laporan teknis yang tersedia di, yang mencakup contoh-contoh realisme luar biasa, yang mewakili gelombang model berikutnya, kali ini menampilkan gambar bergerak.

Meskipun akses yang sangat terbatas ke model seperti Sora dan biaya yang sangat tinggi terkait dengan replikasi model yang serupa dengan pencapaian OpenAI, yang menciptakan kesenjangan signifikan dalam reproduktifitas teknologi, hal ini tetap menunjukkan bahwa batasan AI generatif terus didorong maju. Kekhawatiran utama yang terkait dengan kemajuan ini berkisar pada kesenjangan signifikan dalam infrastruktur komputasi yang diperlukan untuk melatih model-model ini.

Hal ini sangat menghambat pengembangan model terbuka, sehingga terkesan model-model tersebut terutama berfokus pada area yang sangat spesifik, terutama dalam industri kreatif seperti produksi film atau periklanan komersial. Mungkin di tahun-tahun mendatang, kita akan melihat hambatan teknologi untuk mengembangkan teknologi ini berkurang, memungkinkan tersedianya model-model pembangkit video yang lebih ringan yang dapat mengarah pada adopsi teknologi mutakhir ini secara luas. Namun, untuk saat ini, tampaknya hal ini terbatas pada beberapa laboratorium terdepan, yang sebagian besar bersifat komersial.

Dengan begitu banyak peserta dalam persaingan ini, sulit untuk memprediksi apa yang akan terjadi di bidang ini. Sifat skenario yang sangat dinamis ini menyulitkan untuk meramalkan lintasan AI di tahun-tahun mendatang. Memahami bagaimana LLM benar-benar dibandingkan, kekuatan dan kelemahannya, sangat penting saat kita memasuki proses refleksi etis, yang menjadi fokus kesimpulan buku ini.

Menentukan LLM mana yang lebih unggul dalam tugas tertentu atau dalam memahami bahasa dalam domain tertentu merupakan tugas yang menantang yang mengharuskan komunitas memantau secara ketat berbagai inisiatif yang muncul di bidang ini. Pada dasarnya, ada berbagai set data evaluasi yang dirancang untuk menilai berbagai kemampuan LLM, biasanya berdasarkan kuesioner pilihan ganda dengan empat kemungkinan jawaban di berbagai subjek. Seperti yang disebutkan sebelumnya, MMLU adalah salah satu set data kuesioner yang paling banyak digunakan untuk mengevaluasi LLM. Ini mencakup kuesioner di 57 topik yang berbeda. Topik-topik ini menonjolkan disiplin ilmu seperti Ilmu Sosial (12), Humaniora (13), STEM (19), dan lainnya (13).

Disiplin ilmu ini mencakup pengetahuan di berbagai tingkat pendidikan, mulai dari perguruan tinggi, sekolah menengah atas, hingga tingkat universitas. Menjadi set data yang dirancang di Belahan Bumi Utara dan terutama didasarkan pada kuesioner yang terkait dengan tes standar, ada campuran topik universal dan lainnya yang mencerminkan konteks lokal. Di antara yang terakhir, pengetahuan yang terkait dengan humaniora disorot, seperti Sejarah AS,

Sejarah Eropa, Sejarah Dunia, terutama Barat, dan ilmu sosial seperti kebijakan luar negeri AS. Area lain dengan jangkauan yang lebih global terutama mencakup topik STEM.

Penggunaan sumber daya ini dalam evaluasi LLM, meskipun memiliki cakupan universal di beberapa bidang tertentu, memunculkan diskusi tentang penyertaan topik-topik yang relevan secara lokal. Dampaknya terhadap papan peringkat dan, oleh karena itu, insentif bagi pengembang LLM untuk memasukkan topik-topik ini terlihat jelas, yang kami sebut sebagai bias evaluasi di bab-bab sebelumnya buku ini. Kebutuhan akan sumber daya yang mengintegrasikan keragaman budaya yang lebih besar, terutama dalam disiplin ilmu sosial dan humaniora, sangat mendesak karena hal ini mendistorsi lanskap pengembangan dan evaluasi LLM.

Sumber daya lain yang sangat relevan dan banyak digunakan dalam papan peringkat LLM meliputi kumpulan data evaluasi Matematika (MATH Lvl 5), GPQA (pengetahuan lanjutan dalam ilmu dasar), dan MMLU-PRO, versi MMLU yang lebih canggih di mana kuesionernya mencakup 10 alternatif, bukan 4. Aspek yang sangat relevan dari evaluasi LLM berpusat pada kemampuan penalaran mereka. Yang menonjol di antaranya adalah BBH (Big Bench Hard), yang mencakup tugas-tugas terkait penalaran algoritmik dan pengetahuan dunia, dan MuSR, sebuah kumpulan data untuk mengevaluasi penalaran multilangkah (pertanyaan multiturn) yang juga melibatkan masalah algoritmik kompleks. Masalah-masalah tersebut terutama berbasis skenario dan mencakup misteri pembunuhan, pertanyaan penempatan objek, dan optimasi alokasi tim.

Mengenai papan peringkat, terdapat beberapa inisiatif. Salah satu contoh penting adalah Papan Peringkat LLM Terbuka, yang tersedia di Huggingface, yang berfokus pada LLM non-komersial. Selain itu, terdapat inisiatif di mana pengguna merancang kuesioner, dan berdasarkan tanggapannya, poin diberikan untuk berbagai LLM. Dalam konteks ini, papan peringkat Chatbot Arena menonjol, menampilkan LLM terbuka dan komersial. Per 12 Agustus 2024, papan peringkat ini telah mengumpulkan 1.671.145 suara berdasarkan evaluasi 128 model. Pada tanggal tersebut, posisi teratas dalam skor arena (skor dihitung berdasarkan evaluasi pengguna) dipegang oleh ChatGPT-4o (versi 8 Agustus 2024), diikuti oleh Google (DeepMind) Gemini-1.5-Pro-Exp-0801. Posisi ketiga, berdasarkan skor arena, ditempati oleh GPT-40 (versi 13 Mei 2024).

Terakhir, posisi keempat ditempati oleh GPT-4o-mini, versi ringan GPT-4o dalam hal parameter, bersama Claude 3.5 Sonnet oleh Anthropic, Gemini Advanced App oleh Google, dan Llama-3.1-405b-Instruct oleh Meta. Patut dicatat bahwa tidak ada LLM terbuka yang mampu bersaing dengan perusahaan-perusahaan besar di papan peringkat. Menarik juga untuk dicatat bahwa di antara perusahaan-perusahaan yang memimpin persaingan LLM, dua pemain ternama, Google dan Meta, menonjol, sementara dua pendatang baru, OpenAI dan Anthropic, telah muncul, keduanya produk dari era LLM. Ini adalah persaingan yang ketat dengan hasil yang tidak pasti.

Sehubungan dengan kemampuan pemahaman bahasa visual yang dibahas dalam bab ini, papan peringkat yang sama menunjukkan bahwa lebih sedikit model yang berkompetisi. Chatbot Arena menampilkan 15 model yang dievaluasi berdasarkan 65.415 suara (per 11

Agustus 2024). Dua model multimoda, Gemini-1.5-Pro-Exp-0801 dan GPT-4o (versi 13 Mei 2024), memimpin peringkat. Posisi ketiga dipegang oleh Claude 3.5 Sonnet. Secara keseluruhan, dominasi model proprietary pada papan peringkat ini cukup signifikan, dengan hanya beberapa model terbuka yang muncul di peringkat teratas, seperti model InternVL2-26b oleh OpenGVLab, LLaVA-v1.6-34B oleh LLaVa, dan CogVLM2-llama3-chat-19b oleh Zhipu AI. Model-model lainnya semuanya proprietary, menunjukkan persaingan yang ketat antara model Gemini dari Google, GPT dari OpenAI, dan Claude dari Anthropic.

11.9 RISIKO MODEL MULTIMODA

Saat ini, dengan gambaran yang jelas tentang bagaimana LLM, baik berbasis teks percakapan maupun yang terbaru dengan fitur multimoda, telah mendominasi kemajuan paling signifikan dalam AI dalam beberapa tahun terakhir, kita dapat merenungkan risiko yang terlibat dalam pengembangan teknologi ini.

Meskipun kemunculan konten manipulasi mendahului perkembangan AI generatif, terutama melalui teknologi yang melibatkan penyelarasan dan fusi video atau gambar, potensi penyebaran konten manipulasi secara luas seperti deepfake kini sangat besar. Mengenai deepfake, diketahui bahwa hasil paling awal berdasarkan teknologi seperti pertukaran wajah menggunakan algoritma visi komputer konvensional, seperti deteksi titik minat. Proses pembuatan konten semacam itu jauh lebih kompleks, membutuhkan pengembangan dan pengaturan khusus untuk membuatnya.

Dengan kata lain, teknologi aslinya memerlukan pekerjaan penyuntingan yang ekstensif dan pengaturan yang melibatkan, misalnya, perekaman konten untuk meniru suara atau gambar untuk pertukaran wajah. Karena kompleksitas pengaturan eksperimental, kontennya bisa sangat realistis, tetapi menghasilkan konten jenis ini dalam jumlah besar sangatlah menantang. Perubahan yang terjadi dengan AI generatif adalah kemampuan untuk menghasilkan gambar sesuka hati berdasarkan perintah, atau bahkan video, yang secara signifikan meningkatkan potensi membanjiri ekosistem informasi dengan konten ini. Karena hambatan untuk menghasilkan gambar sangat realistis, dan dalam waktu dekat, video, telah dikurangi, akses ke teknologi ini kini lebih luas, dan potensi penyalahgunaannya pun meningkat. Beberapa peneliti, seperti Emilio Ferrara, memperkirakan skenario yang kondusif bagi pencurian identitas dan penyebaran misinformasi yang meluas di jejaring sosial [101]. Kami memiliki kekhawatiran yang sama dan memperkirakan skenario yang sangat kompleks bagi demokrasi Barat dan kebebasan berekspresi. Ini akan menjadi tanggung jawab tidak hanya perusahaan besar, tetapi juga lembaga regulator dan masyarakat sipil untuk memastikan bahwa kemajuan pesat yang dihasilkan oleh AI tidak merugikan kita.

11.10 KESIMPULAN

Model bahasa multimoda merupakan kemajuan signifikan dalam AI, khususnya dalam pembuatan dan manipulasi konten gambar/teks multimoda. Model-model ini telah menunjukkan kemampuan luar biasa dalam menghasilkan teks yang terkadang dapat melampaui kinerja manusia, menunjukkan potensinya untuk merevolusi industri seperti

keuangan, bisnis, layanan kesehatan, dan keamanan siber. Kemampuannya untuk memahami dan menghasilkan konten yang kompleks sangat menjanjikan untuk meningkatkan aplikasi AI di berbagai domain. Namun, penerapan model-model ini bukan tanpa kekhawatiran dan risiko etika.

Model bahasa multimoda secara tidak sengaja dapat menghasilkan konten yang berbahaya, termasuk teks yang menyinggung dan gambar yang tidak pantas, yang menimbulkan risiko etika yang substansial. Tantangan etika lainnya termasuk potensi model-model ini untuk melanggengkan bias dan mengikis pengetahuan kolektif dalam ekosistem digital, yang menyebabkan dilema sosial dan epistemik yang signifikan.

Selain itu, kekhawatiran tentang kredibilitasnya dan potensi risiko keselamatan dan keamanan yang ditimbulkannya menggarisbawahi kebutuhan krusial untuk mengembangkan kerangka kerja yang memprioritaskan keadilan, transparansi, keterjelasan, dan akuntabilitas. Untuk mengatasi kekhawatiran ini, berbagai upaya besar sedang dilakukan untuk mengembangkan mekanisme yang kuat guna memitigasi risiko yang terkait dengan teknologi-teknologi ini. Metrik toksisitas inovatif, metode detoksifikasi, dan algoritma yang menyelaraskan nilai-nilai etika dengan model bahasa saat ini sedang diteliti dan diimplementasikan.

Lebih lanjut, mendorong inovasi yang bertanggung jawab dan menciptakan taksonomi komprehensif atas risiko etika dan sosial merupakan langkah penting menuju integrasi model bahasa multimoda ke dalam bidang-bidang sensitif, termasuk bidang medis. Upaya-upaya ini menggarisbawahi komitmen untuk memastikan bahwa manfaat dari teknologi-teknologi ini terwujud sekaligus meminimalkan potensi bahayanya.

BAB 12
PERSPEKTIF DAN TANTANGAN

12.1 PENDAHULUAN

Di sepanjang buku ini, kami telah mengeksplorasi prinsip-prinsip dasar etika AI, menyajikan contoh-contoh kunci, isu-isu, dan definisi. Tujuan kami adalah untuk menunjukkan tidak hanya pentingnya topik-topik ini tetapi juga tantangan rumit yang terlibat dalam pengembangan etika AI.

Dalam bab terakhir ini, kami akan membahas dua aspek tambahan dari diskusi ini. Pertama, kami bertujuan untuk mengklarifikasi apa yang kami yakini sebagai integrasi etika ke dalam AI, khususnya melalui apa yang kami sebut "pengungkapan etika", yang kami anggap penting bagi kemajuan utama dalam gelombang ketiga etika AI yang sedang berkembang. Kedua, kami akan secara singkat membagikan visi kami untuk masa depan pengembangan AI di tahun-tahun mendatang, mengidentifikasi area-area penting yang perlu diperhatikan, yang menurut pandangan kami, harus dipertimbangkan secara cermat seiring perkembangan bidang ini.

12.2 PENGUNGKAPAN ETIKA UNTUK GELOMBANG KETIGA AI

Sebagaimana disebutkan dalam Bab 3, penelitian terbaru menunjukkan bahwa gelombang ketiga pengembangan AI melibatkan pergeseran ke arah pertimbangan keberlanjutan dan dampak ekologis. Namun, kami juga ingin menekankan tantangan signifikan lainnya terkait kritik yang dibahas di Bab 1, di mana beberapa pihak melabeli prinsip-prinsip etika AI sebagai "tidak berguna".

Meskipun gelombang kedua etika AI berfokus pada peningkatan kesadaran tentang pentingnya pertimbangan etika, tantangan saat ini adalah menerjemahkan kesadaran tersebut ke dalam praktik dan metodologi spesifik yang secara efektif mengintegrasikan etika AI ke dalam fase desain, pengembangan, dan implementasi. Johnson dan Verdicchio telah memberikan panduan konseptual untuk memahami integrasi ini. Mereka mengkritik gagasan sederhana bahwa etika dapat langsung "ditambahkan" ke AI untuk menciptakan apa yang mungkin disebut "AI etis", seperti sistem pembelajaran mesin dengan nilai-nilai etika yang tertanam.

Dalam analisis mereka, mereka menggambarkan keyakinan ini sebagai "kekeliruan penjumlahan", dengan alasan bahwa menggabungkan AI dengan etika tidak secara otomatis menghasilkan sistem etika karena kedua elemen tersebut secara ontologis berbeda. AI harus dipahami sebagai seperangkat artefak komputasional, sementara etika adalah seperangkat nilai-nilai kemanusiaan dan sosial yang tidak dapat begitu saja diintegrasikan ke dalam AI, karena keduanya termasuk dalam kategori ontologis yang berbeda.

Para penulis menjelaskan bahwa agar etika dan AI dapat terintegrasi secara efektif, keduanya harus memiliki karakteristik ontologis yang sama, yang menyiratkan bahwa prinsip-prinsip etika perlu dapat dikomputasi. Namun, mereka berpendapat bahwa hal ini mustahil

karena nilai-nilai etika pada dasarnya abstrak, sosial, dan kontekstual, sementara artefak komputasional berfungsi berdasarkan proses dan data konkret yang deterministik, yang seringkali dapat direduksi menjadi istilah biner dan nilai numerik. Jadi, apa saja pilihan yang kita miliki?

Sebagaimana telah dibahas di seluruh buku ini, kita mengonseptualisasikan AI sebagai sistem sosioteknis. Perspektif ini memungkinkan kita untuk mengeksplorasi cara-cara mengintegrasikan etika ke dalam AI tanpa jatuh ke dalam kekeliruan memperlakukan etika sebagai komponen eksternal yang dapat begitu saja "ditambahkan" ke dalam AI. Ketika AI dipandang semata-mata sebagai artefak komputasional, etika dan AI tidak dapat dengan mudah digabungkan karena keduanya termasuk dalam kategori ontologis yang berbeda. Namun, hubungan antara keduanya menjadi lebih bernuansa dan kompleks ketika seseorang memahami AI sebagai sistem sosioteknis—di mana AI dan nilai-nilai hidup berdampingan dalam kerangka ontologis yang sama.

Dalam kerangka sosioteknis ini, mengintegrasikan nilai-nilai ke dalam AI bukanlah tentang "menanamkan" prinsip-prinsip etika ke dalam artefak teknologi. Sebaliknya, hal itu melibatkan pengakuan bagaimana teknologi dan masyarakat secara kolaboratif membentuk dan memproduksi bersama makna dan nilai-nilai. Perspektif ini menunjukkan bahwa nilai-nilai tidak secara inheren "di dalam" AI tetapi justru dikaitkan dengan teknologi oleh manusia melalui interaksi mereka dan dalam konteks tertentu. Akibatnya, pertimbangan etika dalam pengembangan AI melampaui masalah teknis pengkodean atau desain dan memerlukan pendekatan yang lebih luas yang mempertimbangkan interaksi rumit antara teknologi dan masyarakat.

Dengan mengikuti pendekatan sosioteknis ini, kita dapat menyadari bahwa mencapai pertimbangan etika tersebut memerlukan pendekatan kontekstual terhadap etika dalam AI. Salah satu alternatifnya adalah menekankan pentingnya menerapkan proses hermeneutika yang berkelanjutan. Proses hermeneutika tidak statis; ia berkembang seiring munculnya wawasan dan pemahaman baru. Proses berulang ini berarti bahwa interpretasi terus berlangsung, dengan setiap lapisan pemahaman baru memengaruhi dan membentuk kembali interpretasi sebelumnya. Ini mencakup konteks historis, budaya, dan sosial, dengan mengakui bahwa faktor-faktor ini memengaruhi makna dan tidak dapat sepenuhnya dipahami secara terpisah. Lebih lanjut, dalam proses hermeneutik, penafsir harus menyadari bias, prakonsepsi, dan asumsi mereka sendiri, sehingga memeriksa dan, jika perlu, menantang bias ini untuk memungkinkan pemahaman yang lebih autentik dan mendalam.

Oleh karena itu, dengan memandang AI sebagai sistem sosioteknis, pengungkapan etika membuka dimensi konvergensi baru, di mana AI dan etika tidak dilihat sebagai domain yang terpisah tetapi sebagai komponen yang saling berhubungan dari kerangka kerja masyarakat yang lebih luas. Dalam pandangan ini, teknologi AI bukan hanya alat tetapi bagian dari jaringan kompleks hubungan manusia, norma sosial, dan praktik organisasi. Pengungkapan etika, seperti yang disajikan oleh Arriagada-Bruneau, terlibat dengan sikap filosofis, konsep Heidegger tentang "Gestell" atau pembingkaian teknologi, yang menggambarkan bagaimana teknologi modern membentuk hubungan kita dengan dunia,

mereduksi segalanya menjadi sumber daya yang tersedia untuk dieksploitasi. Menurut Heidegger, kerangka ini membatasi pemahaman kita tentang esensi sejati teknologi, membatasinya pada nilai instrumentalnya.

Dalam konteks AI, kerangka teknologi ini dapat bermanifestasi sebagai "tekno-chauvinisme" atau "solusionisme teknologi", di mana teknologi dipandang sebagai solusi utama bagi permasalahan sosial yang kompleks, seringkali mengabaikan dimensi etika, politik, dan sosial yang mendasarinya. Sebuah pengungkapan etika berupaya membebaskan AI dari kerangka teknologi ini, mendorong hubungan dengan teknologi yang melampaui fungsi utilitariannya. Pendekatan ini menantang gagasan bahwa nilai AI semata-mata terletak pada kapasitasnya untuk mengoptimalkan proses dan membuat keputusan. Sebaliknya, pendekatan ini menganjurkan AI sebagai teknologi yang dapat memperkaya pemahaman kita tentang dunia dan tempat kita di dalamnya.

Implikasi praktis dari pengungkapan etika sangatlah mendalam. Hal ini membutuhkan integrasi pertimbangan etika sejak awal pengembangan AI, memastikan bahwa etika bukanlah renungan belakangan, melainkan komponen fundamental dari proses desain. Integrasi ini menuntut dialog berkelanjutan antara pengembang, ahli etika, pembuat kebijakan, dan masyarakat luas, yang mendorong pemahaman bersama tentang tantangan dan peluang etika AI. Lebih lanjut, pengungkapan etika menuntut pendekatan yang dinamis dan adaptif terhadap Etika AI, di mana pedoman etika berkembang sebagai respons terhadap informasi, perspektif, dan perubahan sosial baru. Pendekatan ini berbeda dengan penerapan prinsip-prinsip etika yang statis, yang menekankan perlunya refleksi dan revisi etika yang berkelanjutan di sepanjang siklus hidup sistem AI. Dengan merangkul proses hermeneutik, etika AI dapat beralih dari reaktif dan spesifik situasi menjadi proaktif dan integral dengan inovasi teknologi.

Untuk menerapkan konsep pengungkapan etika sebagai proses hermeneutik secara efektif, seseorang harus mengadopsi pendekatan yang dinamis dan berlandaskan konteks yang mengintegrasikan etika sebagai komponen intrinsik sejak awal proyek AI. Proses ini dimulai dengan analisis kontekstual yang komprehensif, dengan mengakui dimensi historis, budaya, dan sosial yang membentuk lanskap etika tempat sistem AI akan beroperasi dan asumsi atau batasan apa pun yang diakui oleh tim pengembang. Pengungkapan etika harus menantang paradigma teknosentris tradisional di mana kita menggunakan lebih banyak teknologi untuk memperbaiki teknologi dan, sebagai gantinya, menanamkan pertimbangan etika ke dalam fase desain, memastikan bahwa tujuan dan operasi AI selaras dengan nilai-nilai sosial, alih-alih sekadar tujuan teknis.

Integrasi ini menuntut refleksi hermeneutik berkelanjutan (mendalam dan berulang) di seluruh siklus hidup AI dan ekosistem di sekitarnya, di mana pengembang, perancang, pengguna, dan pemangku kepentingan lainnya terlibat dalam proses interpretasi dan reinterpretasi yang berulang dan berkelanjutan, secara aktif mempertanyakan dan menyempurnakan bias dan asumsi mereka seiring perkembangan proyek. Lebih lanjut, keterlibatan para pemangku kepentingan ini krusial untuk memastikan beragam perspektif menginformasikan kerangka etika.

Pendekatan ini mengharuskan pengembangan pedoman etika dinamis yang tetap adaptif terhadap wawasan baru dan perubahan sosial. Selain itu, mekanisme pemantauan dan umpan balik yang berkelanjutan harus ditetapkan untuk menilai dan menyelaraskan kembali kinerja AI dengan komitmen etikanya. Melalui langkah-langkah ini, pengungkapan etika sebagai proses hermeneutik dapat diterapkan pada metodologi dan praktik yang ada dalam bidang AI, memastikan bahwa pengembangan AI dipandu oleh fondasi etika yang kuat dan peka konteks.

12.3 PRAKATA MENUJU MASA DEPAN AI

Seiring AI terus berkembang pesat dan diterapkan di berbagai bidang, kami menguraikan ekspektasi kami untuk perkembangan etika AI dan teknologi informasi di masa depan, melampaui kemajuan yang dibutuhkan dalam keberlanjutan dan reproduktifitas yang dibahas dalam Bab 3.

Kami percaya bahwa AI harus dianggap sebagai revolusi teknologi lain dalam kontinum inovasi manusia yang panjang. Seperti transformasi teknologi besar sebelumnya, AI menghadirkan peluang dan tantangan bagi perekonomian dan budaya kita. Sejarah menunjukkan kepada kita bahwa teknologi menantang masyarakat dan mendorong proses transformatif yang mendalam. Transformasi melibatkan perubahan dalam pendidikan, dalam cara kita berhubungan dengan pekerjaan kita, dalam cara kita berinteraksi dengan orang lain, dan, tentu saja, dalam banyak aspek kehidupan kita sehari-hari yang rentan terhadap teknologi disruptif. Esensi perubahan terletak pada kemampuan adaptasi institusi di tingkat makro, tetapi yang terpenting, pada kemampuan adaptasi individu.

Buku ini menunjukkan bahwa kita sedang mengalami perubahan teknologi yang mendalam yang didorong oleh AI. Meskipun sistem cerdas telah menjadi bagian dari kehidupan kita selama bertahun-tahun, mencakup beragam aktivitas manusia, yang menjadi ciri interaksi ini adalah ketidaksadaran akan kehadiran AI. Dalam interaksi kita dengan sistem informasi, AI telah membantu kita menemukan film atau memesan tiket pesawat. Namun, AI juga beroperasi di sisi belakang (back-end) berbagai layanan yang kita gunakan sehari-hari, seperti pemrosesan citra medis, deteksi penipuan di perbankan, dan deteksi ancaman keamanan siber. Namun, kehadiran ini sebagian besar tidak terlihat. Kita telah hidup berdampingan dengan AI tanpa sepenuhnya menyadari interaksi kita dengannya.

Perubahan teknologi yang kita alami membawa kita ke skenario baru. Kehadiran AI kini lebih diakui. Kita kini tahu bahwa AI-lah yang membuat Alexa berinteraksi dengan kita. Kita kini memahami bahwa saat menggunakan ChatGPT, kita berinteraksi dengan AI. Kita menjadi sadar akan interaksi kita dengan AI, yang disebabkan oleh AI yang berpindah dari sisi belakang sistem informasi ke sisi depan, berinteraksi langsung dengan kita dalam sistem percakapan, atau membantu kita dalam tugas sehari-hari. Pengaruhnya kini nyata, dan karenanya jauh lebih mendalam. Studi tentang interaksi antara AI dan manusia kemungkinan akan menjadi salah satu bidang yang paling pesat perkembangannya di tahun-tahun mendatang. Lebih lanjut, kemampuan AI untuk meniru manusia mungkin merupakan jalur paling menakjubkan yang

terbentang di masa depan bagi kita. Dengan sistem multi-agen yang didukung oleh LLM, kita dapat mensimulasikan interaksi manusia.

Kemungkinan untuk mengantisipasi, melalui AI, apa yang diharapkan dari interaksi manusia dapat menjadi sangat menarik. Teknologi yang kita miliki memungkinkan kita untuk mensimulasikan rapat. Simulasi memungkinkan adanya kontrol, dan oleh karena itu, dalam skenario ini, menjadi lebih memungkinkan untuk meramalkan faktor-faktor penting yang memengaruhi hasil interaksi manusia. Kita akan semakin sering mendengar suara-suara yang menganjurkan penggantian pembuat kebijakan dengan agen cerdas, dan pembuat keputusan dengan AI, semuanya untuk mengurangi dampak buruk yang melekat dalam interaksi manusia, seperti perebutan pengaruh dan, pada akhirnya, korupsi. Dapatkah AI membantu kita membangun demokrasi yang lebih tepercaya dan transparan? Ini adalah perdebatan yang niscaya akan muncul di tahun-tahun mendatang.

Kekhawatiran yang signifikan muncul terkait lambatnya adaptasi kelembagaan terhadap perubahan teknologi. Ini adalah dua proses yang beroperasi dalam dinamika yang sangat berbeda. Meskipun perkembangan teknologi mengalami proses perubahan yang cepat, negara-negara bergulat dengan demokrasi liberal yang dicirikan oleh kelemahan kelembagaan yang meluas dan meningkatnya ketidakpercayaan terhadap lembaga-lembaga tersebut. Isu utama adalah seberapa kuat, kredibel, dan andal penerapan kerangka regulasi dalam skenario pelemahan kelembagaan yang progresif.

Lambatnya adaptasi kelembagaan menghadirkan skenario yang kompleks. Tampaknya kecepatan individu mengadopsi teknologi jauh melampaui cara lembaga mengadaptasi regulasi mereka untuk memastikan penggunaan AI yang aman. Perdebatan sengit seputar peran multilateralisme dan kebutuhan untuk menangani ruang terkoordinasi bagi tindakan regulasi di tingkat global sangatlah penting. Masih belum jelas apa hasil dari upaya ini, dengan inisiatif yang datang dari beragam kerangka kelembagaan yang melibatkan entitas mulai dari PBB hingga OECD atau UNESCO. Satu-satunya kepastian adalah bahwa kerangka regulasi ini akan membutuhkan upaya yang luar biasa untuk terus memperbaruinya.

Upaya lain juga diperlukan. Kami mengharapkan kemajuan signifikan dalam membangun infrastruktur sosioteknis yang dirancang untuk mengatasi beberapa tantangan etika yang paling mendesak. Bagi pengembang AI, tampaknya regulasi yang muncul dan ekspektasi masyarakat akan menyiratkan pengembangan standar baru, termasuk kode etik dan sertifikasi. Hal ini juga akan mengarah pada praktik-praktik baru yang membutuhkan pelatihan yang lebih baik untuk mengatasi masalah etika dan, yang terpenting, pembentukan tim interdisipliner untuk memandu pengembangan AI.

Lebih lanjut, seiring dengan semakin banyaknya regulasi yang mewajibkan transparansi dan kemudahan dijelaskan, jenis-jenis profesional dan organisasi baru, yang berpotensi saling terhubung secara longgar, kemungkinan akan muncul untuk mewakili kebutuhan warga negara dan mengadvokasi jenis-jenis AI yang dapat (atau tidak dapat) diterapkan serta tingkat transparansi dan akuntabilitas yang dibutuhkan di setiap sektor masyarakat kita. Untuk mengatasi tantangan terkait ketenagakerjaan, kami mengantisipasi peningkatan penelitian berbasis bukti yang mengkaji dampak AI terhadap pasar kerja.

Meskipun dampak AI diproyeksikan signifikan, kami yakin bahwa transisi dari studi prediktif ke studi observasional akan memberikan pemahaman yang lebih jelas tentang aspek-aspek rumit pekerjaan yang tidak dapat sepenuhnya diotomatisasi atau digantikan oleh AI. Akibatnya, kami memperkirakan akan terjadi pergeseran fokus dari merancang AI menjadi menggantikan peran manusia, dan menciptakan konfigurasi kerja manusia-AI yang lebih efektif dan bermanfaat. Dengan demikian, tantangan yang kita hadapi dalam menyelaraskan AI dengan nilai-nilai kemanusiaan sangat besar, tetapi bukan berarti mustahil untuk diatasi. Transformasi yang dibawa AI kepada masyarakat merupakan peluang untuk menata kembali bagaimana teknologi dapat melayani kebaikan bersama, mendorong masa depan yang lebih transparan, setara, dan adil.

Potensi AI yang sesungguhnya tidak hanya terletak pada kapasitasnya untuk mengotomatiskan tugas atau menghasilkan pengetahuan, tetapi juga pada kemampuannya untuk meningkatkan agensi dan pengambilan keputusan manusia. Dengan memastikan bahwa prinsip-prinsip etika tertanam dalam setiap aspek ekosistem AI, kita dapat menavigasi revolusi teknologi ini dengan cara yang meningkatkan, alih-alih mengurangi, kemanusiaan kita. Merupakan tanggung jawab kolektif kita untuk membentuk AI menjadi kekuatan untuk perubahan yang positif dan transformatif.

Mengintegrasikan etika ke dalam AI membutuhkan evolusi interdisipliner yang mendalam di tahun-tahun mendatang. Dalam buku ini, kami menegaskan bahwa integrasi ini jauh melampaui sekadar menambahkan perangkat etika ke dalam perangkat AI. Integrasi ini menuntut perubahan mendasar dan tegas dalam cara kita memahami dan mengembangkan AI. Dengan mengakui sistem AI sebagai konstruksi sosioteknis, integrasi etika menuntut pendekatan holistik yang membahas dimensi teknologi dan sosial AI. Kami berharap buku ini dapat membantu menyampaikan urgensi transformasi ini, menginspirasi kita semua untuk menjadi agen perubahan yang aktif dalam membentuk masa depan AI yang lebih etis dan bertanggung jawab.

DAFTAR PUSTAKA

Anderson, S. L., & Anderson, M. (2021). AI and ethics. *AI and Ethics*, *1*(1), 27-31.

Bartneck, C., Lütge, C., Wagner, A., & Welsh, S. (2021). *An introduction to ethics in robotics and AI* (p. 117). Springer Nature.

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.

Borenstein, J., & Howard, A. (2021). Emerging challenges in AI and the need for AI ethics education. *AI and Ethics*, *1*(1), 61-65.

Braun, M., Hummel, P., Beck, S., & Dabrock, P. (2021). Primer on an ethics of AI-based decision support systems in the clinic. *Journal of medical ethics*, *47*(12), e3-e3.

Brey, P., & Dainow, B. (2024). Ethics by design for artificial intelligence. *AI and Ethics*, *4*(4), 1265-1277.

Brossi, L., Castillo, A. M., & Cortesi, S. (2022). Student-centred requirements for the ethics of AI in education. In *The ethics of artificial intelligence in education* (pp. 91-112). Routledge.

Caliskan, A. (2023, August). Artificial Intelligence, Bias, and Ethics. In *IJCAI* (pp. 7007-7013).

Canca, C. (2020). Operationalizing AI ethics principles. *Communications of the ACM*, *63*(12), 18-21.

Carrillo, M. R. (2020). Artificial intelligence: From ethics to law. *Telecommunications policy*, *44*(6), 101937.

Chauhan, C., & Gullapalli, R. R. (2025). Ethics of AI in pathology: current paradigms and emerging issues. *Artificial Intelligence in Pathology*, 159-180.

Chinta, S. V., Wang, Z., Yin, Z., Hoang, N., Gonzalez, M., Quy, T. L., & Zhang, W. (2024). FairAIED: Navigating fairness, bias, and ethics in educational AI applications. *arXiv preprint arXiv:2407.18745*.

Choung, H., David, P., & Ross, A. (2023). Trust and ethics in AI. *Ai & Society*, *38*(2), 733-745.

Cox, A. (2022). The ethics of AI for information professionals: Eight scenarios. *Journal of the Australian Library and Information Association*, *71*(3), 201-214.

Dubber, M. D., Pasquale, F., & Das, S. (Eds.). (2020). *The Oxford handbook of ethics of AI*. Oxford University Press.

Elmahjub, E. (2023). Artificial intelligence (AI) in Islamic ethics: Towards pluralist ethical benchmarking for AI. *Philosophy & Technology*, *36*(4), 73.

Furey, H., & Martin, F. (2019). AI education matters: a modular approach to AI ethics education. *AI Matters*, *4*(4), 13-15.

Garrett, N., Beard, N., & Fiesler, C. (2020, February). More than" If Time Allows" the role of ethics in AI education. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 272-278).

Gerdes, A. (2022). The tech industry hijacking of the AI ethics research agenda and why we should reclaim it. *Discover Artificial Intelligence*, *2*(1), 25.

Goldsmith, J., & Burton, E. (2017, February). Why teaching ethics to AI practitioners is important. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 31, No. 1).

Hagendorff, T. (2022). Blind spots in AI ethics. *AI and Ethics*, *2*(4), 851-867.

Heilinger, J. C. (2022). The ethics of AI ethics. A constructive critique. *Philosophy & Technology*, *35*(3), 61.

Holmes, W., Iniesto, F., Anastopoulou, S., & Boticario, J. G. (2023). Stakeholder perspectives on the ethics of AI in distance-based higher education. *International Review of Research in Open and Distributed Learning*, *24*(2), 96-117.

Huang, C., Zhang, Z., Mao, B., & Yao, X. (2022). An overview of artificial intelligence ethics. *IEEE Transactions on Artificial Intelligence*, *4*(4), 799-819.

Hunkenschroer, A. L., & Luetge, C. (2022). Ethics of AI-enabled recruiting and selection: A review and research agenda. *Journal of Business Ethics*, *178*(4), 977-1007.

Kang, Y., Zhang, Q., & Roth, R. (2023). The ethics of AI-Generated maps: A study of DALLE 2 and implications for cartography. *arXiv preprint arXiv:2304.10743*.

Kazim, E., & Koshiyama, A. S. (2021). A high-level overview of AI ethics. *Patterns*, *2*(9).

Kenig, N., Monton Echeverria, J., & Rubi, C. (2024). Ethics for AI in plastic surgery: guidelines and review. *Aesthetic Plastic Surgery*, *48*(11), 2204-2209.

Khan, A. A., Badshah, S., Liang, P., Waseem, M., Khan, B., Ahmad, A., ... & Akbar, M. A. (2022, June). Ethics of AI: A systematic literature review of principles and challenges.

In *Proceedings of the 26th international conference on evaluation and assessment in software engineering* (pp. 383-392).

Kiemde, S. M. A., & Kora, A. D. (2022). Towards an ethics of AI in Africa: rule of education. *AI and Ethics*, *2*(1), 35-40.

Lauer, D. (2021). You cannot have AI ethics without ethics. *AI and Ethics*, *1*(1), 21-25.

Li, F., Ruijs, N., & Lu, Y. (2022). Ethics & AI: A systematic review on ethical concerns and related strategies for designing with AI in healthcare. *Ai*, *4*(1), 28-53.

Löhr, G. (2024). If conceptual engineering is a new method in the ethics of AI, what method is it exactly?. *AI and Ethics*, *4*(2), 575-585.

McLennan, S., Fiske, A., Celi, L. A., Müller, R., Harder, J., Ritt, K., ... & Buyx, A. (2020). An embedded ethics approach for AI development. *Nature Machine Intelligence*, *2*(9), 488-490.

Melhart, D., Togelius, J., Mikkelsen, B., Holmgård, C., & Yannakakis, G. N. (2023). The ethics of AI in games. *IEEE Transactions on Affective Computing*, *15*(1), 79-92.

Mökander, J., & Floridi, L. (2021). Ethics-based auditing to develop trustworthy AI. *Minds and Machines*, *31*(2), 323-327.

Morley, J., & Floridi, L. (2025). The ethics of AI in healthcare: An updated mapping review. *Ethics and Medical Technology: Essays on Artificial Intelligence, Enhancement, Privacy, and Justice*, 29-57.

Murikah, W., Nthenge, J. K., & Musyoka, F. M. (2024). Bias and ethics of AI systems applied in auditing-A systematic review. *Scientific African*, *25*, e02281.

Mutashar, M. K. (2024). Navigating ethics in AI-driven translation for a human-centric future. *Academia Open*, *9*(2), 10-21070.

Naik, T., Gostu, H., & Sharma, R. (2024, March). Navigating ethics of AI-powered creativity in Midjourney. In *2024 3rd International Conference for Innovation in Technology (INOCON)* (pp. 1-6). IEEE.

Ortega-Bolaños, R., Bernal-Salcedo, J., Germán Ortiz, M., Galeano Sarmiento, J., Ruz, G. A., & Tabares-Soto, R. (2024). Applying the ethics of AI: a systematic review of tools for developing and assessing AI-based systems. *Artificial Intelligence Review*, *57*(5), 110.

Pant, A., Hoda, R., Spiegler, S. V., Tantithamthavorn, C., & Turhan, B. (2024). Ethics in the age of AI: An analysis of AI practitioners' awareness and challenges. *ACM Transactions on Software Engineering and Methodology*, *33*(3), 1-35.

Piedra, J. (2023). Decolonizing the ethics of AI. *Cosmos and History: The Journal of Natural and Social Philosophy*, *19*(1), 467-480.

Porayska-Pomsta, K., Holmes, W., & Nemorin, S. (2023). The ethics of AI in education. In *Handbook of artificial intelligence in education* (pp. 571-604). Edward Elgar Publishing.

Ratti, E., & Graves, M. (2025). A capability approach to AI Ethics. *American Philosophical Quarterly*, *62*(1), 1-16.

Rességuier, A., & Rodrigues, R. (2020). AI ethics should not remain toothless! A call to bring back the teeth of ethics. *Big Data & Society*, *7*(2), 2053951720942541.

Ryan, M. (2020). In AI we trust: ethics, artificial intelligence, and reliability. *Science and Engineering Ethics*, *26*(5), 2749-2767.

Sam, A. K., & Olbrich, P. (2023). The need for AI ethics in higher education. In *AI ethics in higher education: Insights from Africa and beyond* (pp. 3-10). Cham: Springer International Publishing.

Sharma, V., Mishra, N., Kukreja, V., Alkhayyat, A., & Elngar, A. A. (2023, March). Framework for evaluating ethics in AI. In *2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA)* (pp. 307-312). IEEE.

Shih, P. K., Lin, C. H., Wu, L. Y., & Yu, C. C. (2021). Learning ethics in AI—teaching non-engineering undergraduates through situated learning. *Sustainability*, *13*(7), 3718.

Siau, K., & Wang, W. (2020). Artificial intelligence (AI) ethics: ethics of AI and ethical AI. *Journal of Database Management (JDM)*, *31*(2), 74-87.

Singer, P., & Tse, Y. F. (2023). AI ethics: The case for including animals. *AI and Ethics*, *3*(2), 539-551.

Singhal, A., Neveditsin, N., Tanveer, H., & Mago, V. (2024). Toward fairness, accountability, transparency, and ethics in AI for social media and health care: scoping review. *JMIR Medical Informatics*, *12*(1), e50048.

Stahl, B. C. (2021). Concepts of ethics and their application to AI. In *Artificial Intelligence for a better future: an ecosystem perspective on the Ethics of AI and emerging Digital Technologies* (pp. 19-33). Cham: Springer International Publishing.

Timmers, P. (2019). Ethics of AI and cybersecurity when sovereignty is at stake. *Minds and Machines*, *29*(4), 635-645.

Trotta, A., Ziosi, M., & Lomonaco, V. (2023). The future of ethics in AI: challenges and opportunities. *AI & SOCIETY*, *38*(2), 439-441.

Vakkuri, V., Kemell, K. K., & Abrahamsson, P. (2019, November). Implementing ethics in AI: initial results of an industrial multiple case study. In *International Conference on Product-Focused Software Process Improvement* (pp. 331-338). Cham: Springer International Publishing.

Vargas-Murillo, A. R., Pari-Bedoya, I. N. M. D. L. A., & Guevara-Soto, F. D. J. (2023, June). The ethics of AI assisted learning: A systematic literature review on the impacts of ChatGPT usage in education. In *Proceedings of the 2023 8th International Conference on Distance Education and Learning* (pp. 8-13).

Whittlestone, J., & Clarke, S. (2022). AI challenges for society and ethics. In *The Oxford Handbook of AI Governance* (pp. 45-64). New York, NY: Oxford University Press.

Zohny, H., McMillan, J., & King, M. (2023). Ethics of generative AI. *Journal of medical ethics*, *49*(2), 79-80.

ETIKA PADA AI (Artificial Intelligence) dan TI (Teknologi Informasi)

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

BIODATA PENULIS

Dr. Joseph Teguh Santoso, M.Kom memiliki Jabatan Akademik Lektor Kepala dan praktisi industri yang berpengalaman. Saat ini menjabat sebagai Rektor Universitas Sains dan Teknologi Komputer (Universitas STEKOM), salah satu universitas terkemuka di Jawa Tengah, Indonesia. Dengan pengalaman lebih dari 13 tahun di dunia bisnis dan praktisi industri di China, beliau membawa perspektif global dan inovasi yang signifikan ke dalam dunia akademis.Sebagai seorang entrepreneur, penulis adalah pencipta TopLoker.com, sebuah platform inovatif yang merevolusi cara mencari dan menawarkan pekerjaan. TopLoker.com adalah portal lowongan bursa kerja terbesar di Indonesia, khusus untuk pendidikan SMA/SMK sederajat. TopLoker.com telah mendapatkan penghargaan sebagai juara 1 Startup4industry 2022 oleh Kementerian Perindustrian Republik Indonesia. Kontribusi Dr. Joseph dalam menyediakan akses pekerjaan yang luas bagi lulusan SMA/SMK telah membantu banyak individu menemukan peluang kerja yang sesuai dengan keahlian mereka. Selain itu, Dr. Joseph Teguh Santoso, M.Kom adalah pendiri dari dua organisasi yaitu (1) organisasi guru/pendidik PTIC (Perkumpulan Teacherpreneur Indonesia Cerdas) yang bertujuan untuk meningkatkan kualitas pendidikan dan kesejahteraan guru/pendidik dengan wawasan entrepreneurship, serta (2) organisasi industri PERKIVI (Perkumpulan Komunitas Industri dan Vokasi Indonesia) yang berfokus pada pengembangan link and match antara industri dan dunia pendidikan. Sebagai Rektor, Dr. Joseph Teguh Santoso, M.Kom memiliki kepemimpinan yang berorientasi pada hasil, dan berkomitmen untuk mendorong kemajuan Universitas Sains dan Teknologi Komputer (Universitas STEKOM). Saat ini Universitas STEKOM telah mengalami transformasi positif dalam peningkatan kualitas pendidikan, perluasan fasilitas, serta penguatan kemitraan Perguruan Tinggi Nasional dan Internasional. Beliau memprioritaskan pengembangan sumber daya manusia dan penelitian, serta memastikan bahwa universitas berada di garis depan dalam inovasi dan teknologi untuk mencapai tujuan akhir, yaitu lulusan yang mampu bekerja dan sukses setelah lulus. Dr. Joseph Teguh Santoso, M.Kom sering diundang sebagai pembicara di berbagai konferensi nasional maupun internasional dan telah menerima berbagai penghargaan atas dedikasinya dalam bidang pendidikan, industri, dan kewirausahaan.

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 978-634-7227-70-6 (PDF)
9 786347 227706