YAYASAN PRIMA AGUS TEKNIK

TATA KELOLA PROFESIONAL dengan AI (Kecerdasan Buatan)

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

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

TATA KELOLA PROFESIONAL dengan AI (Kecerdasan Buatan)

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 978-634-7227-90-4 (PDF)
9 786347 227904

TATA KELOLA PROFESIONAL dengan AI (Kecerdasan Buatan)

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

ISBN : 978-634-7227-90-4 (PDF)

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

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

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

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

Anggota IKAPI No: 279 / ALB / JTE / 2023

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

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

KATA PENGANTAR

Revolusi kecerdasan buatan (AI) telah mengubah lanskap bisnis, pemerintahan, dan kehidupan sehari-hari dengan kecepatan yang belum pernah terjadi sebelumnya. Dari chatbot cerdas yang melayani pelanggan hingga algoritma prediktif yang mengoptimalkan rantai pasok, AI menjanjikan efisiensi luar biasa. Namun, di balik potensi itu tersembunyi risiko serius: bias diskriminatif, pelanggaran privasi data, kerentanan keamanan siber, hingga dampak etis yang bisa merusak kepercayaan publik. Di sinilah tata kelola profesional AI menjadi krusial, bukan sekadar kewajiban regulasi, melainkan fondasi untuk inovasi yang berkelanjutan dan bertanggung jawab.

Buku “Tata Kelola Profesional dengan AI (Kecerdasan Buatan)” lahir dari kebutuhan mendesak itu. Ditulis dengan perspektif praktis dan berbasis bukti, buku ini menyajikan panduan lengkap bagi eksekutif, manajer IT, pengembang, dan regulator yang ingin mengintegrasikan AI secara aman di organisasi mereka. Khususnya di Indonesia, di mana adopsi AI melonjak seiring transformasi digital nasional—seperti inisiatif Making Indonesia 4.0—buku ini relevan untuk menyelaraskan praktik lokal dengan standar global, termasuk persiapan menghadapi regulasi seperti RUU AI yang sedang digodok.

Struktur buku dirancang secara sistematis untuk membangun pemahaman bertahap. Bab 1 memperkenalkan jenis-jenis AI, risiko potensial seperti halusinasi dan ketergantungan black-box, serta prinsip AI bertanggung jawab yang menjadi landasan utama. Bab 2 membahas kesiapan organisasi, termasuk peran pemangku kepentingan, kolaborasi lintas fungsi, dan pelatihan terminologi AI. Bab 3 fokus pada pembaruan kebijakan untuk pengawasan otonom, privasi data, dan risiko pihak ketiga.

Lebih lanjut, Bab 4 dan 5 mendalami hukum privasi serta perlindungan data—dari pemberitahuan persetujuan hingga kewajiban pengontrol data—serta hukum sektoral seperti kekayaan intelektual, anti-diskriminasi, perlindungan konsumen, dan tanggung jawab produk. Bab 6 secara khusus menguraikan Undang-Undang AI Uni Eropa sebagai benchmark global, dengan klasifikasi risiko dan strategi implementasi. Bab 7 mengeksplorasi standar internasional, mulai dari Prinsip AI OECD, Kerangka Risiko NIST (AI RMF), hingga ISO dan ARIA untuk evaluasi model.

Bagian akhir buku menawarkan blueprint operasional: Bab 8 tentang desain sistem AI berbasis konteks bisnis dan penilaian dampak; Bab 9 pada tata kelola data, silsilah data, dan pengujian; Bab 10 untuk penyebaran, implementasi, dan pemantauan berkelanjutan; Bab 11 mengenai manajemen risiko, audit, dan penanganan insiden; serta Bab 12 yang menutup dengan operasi AI berkelanjutan, termasuk manajemen catatan, komunikasi eksternal, dan penghentian sistem.

Setiap bab dilengkapi ringkasan, contoh kasus nyata, dan rekomendasi actionable, sehingga pembaca tidak hanya memahami "apa" dan "mengapa", tetapi juga "bagaimana"

menerapkannya. Misalnya, bagaimana melakukan penilaian dampak AI untuk meminimalkan bias, atau menyusun kebijakan privasi yang sesuai PDPA Indonesia dan GDPR.

Saya mengundang Anda untuk merefleksikan: Apakah organisasi Anda siap menghadapi era AI yang matang? Buku ini bukan hanya referensi, melainkan alat untuk membangun budaya tata kelola yang proaktif. Dengan demikian, kita dapat mewujudkan AI yang tidak hanya pintar, tapi juga adil, aman, dan bermanfaat bagi masyarakat.

Terima kasih kepada para kontributor, pakar, dan pembaca setia yang membuat karya ini mungkin. Semoga buku ini menjadi penerang di tengah kabut regulasi AI yang semakin kompleks.

Selamat & Semangat Membaca...!!!

Semarang, Februari 2026
Penulis

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

DAFTAR ISI

BAB 1
AI DAN TATA KELOLA AI

1.1 JENIS-JENIS AI

Kecerdasan Buatan (AI) adalah bidang interdisipliner yang menggabungkan ilmu komputer, statistik, matematika, dan teknik untuk menciptakan sistem yang mampu melakukan tugas-tugas yang biasanya membutuhkan kecerdasan manusia. Tujuan utama AI adalah merancang mesin yang dapat berpikir, belajar, dan memecahkan masalah secara mandiri.

Bagian ini mencakup definisi dan contoh berbagai jenis AI.

Definisi AI

Kecerdasan Buatan (AI) adalah simulasi proses kecerdasan manusia oleh mesin, umumnya sistem komputer. Proses-proses ini meliputi pembelajaran, penalaran, pemecahan masalah, persepsi, dan pemahaman bahasa.

Beberapa sistem AI dirancang untuk melakukan tugas-tugas spesifik, seperti pengenalan dan pembuatan gambar, pemrosesan ucapan, pemahaman bahasa alami, pengambilan keputusan, dan banyak lagi. Misalnya, alat seperti HeyGen membuat video realistis dari subjek tertentu dari foto dan skrip.

Sistem AI lainnya dirancang untuk melakukan berbagai tugas, termasuk model bahasa besar (LLM) seperti Claude dari Anthropic dan Microsoft Copilot. Misalnya, asisten virtual seperti Siri dan Alexa menggunakan AI untuk menafsirkan bahasa lisan, memahami permintaan, dan memberikan respons yang relevan.

Kemampuan AI

AI sering dikategorikan ke dalam tingkat kemampuan, berdasarkan sejauh mana kemampuan sistem untuk meniru kemampuan kognitif manusia. Ini adalah:

- AI Sempit (atau AI Lemah) mengacu pada kelas sistem AI yang dirancang dan dilatih untuk melakukan serangkaian tugas spesifik atau sempit. Sistem ini tidak memiliki kecerdasan atau kesadaran umum dan hanya berfokus pada satu jenis fungsi. Salah satu contohnya adalah IBM Deep Blue, AI yang dirancang untuk bermain catur pada tingkat ahli. Contoh lainnya adalah AI model bahasa besar (LLM) (juga dikenal sebagai AI generatif), seperti yang diterapkan dalam ChatGPT, Claude, Grok, dan lainnya.
- Agen AI adalah sistem otonom yang memahami lingkungannya melalui informasi sensor yang masuk dan kemudian mengambil tindakan untuk mencapai tujuan tertentu.
- AI Umum (atau Kecerdasan Buatan Umum—AGI) adalah bentuk AI yang dapat memahami, belajar, dan menerapkan kecerdasan di berbagai tugas, seperti halnya manusia. AGI mampu menyelesaikan hampir semua masalah yang dihadapinya, sama seperti manusia. Misalnya, sistem AI yang mampu mengelola seluruh atau sebagian organisasi, termasuk melakukan tugas-tugas seperti pengambilan keputusan strategis

dan layanan pelanggan, dan bahkan tugas-tugas kreatif seperti menulis atau membuat desain produk atau proses.

- Kecerdasan Buatan Super (ASI) (atau AI Supercerdas) adalah AI yang melampaui kecerdasan manusia dalam semua aspek kreativitas, pemecahan masalah, pemahaman emosional, dan pengambilan keputusan. Meskipun tingkat AI ini belum ada, hal ini sering dibahas dalam penelitian AI dan diskusi etika, karena potensi risiko dan manfaatnya.
- AI yang menyerupai Tuhan adalah AI pada tingkat di luar Kecerdasan Buatan Super (ASI), di mana kemampuan AI melampaui kognisi manusia sedemikian rupa sehingga menjadi tidak dapat dipahami oleh manusia di semua bidang.

Fungsionalitas AI

Sistem AI juga dikategorikan berdasarkan fungsionalitasnya, berdasarkan jenis tugas atau proses yang dapat mereka lakukan. Ini termasuk:

- Mesin reaktif adalah sistem AI yang dirancang untuk merespons rangsangan atau masukan tertentu. Mereka tidak menyimpan ingatan tentang interaksi masa lalu, dan juga tidak belajar dari interaksi tersebut. Mesin reaktif beroperasi berdasarkan aturan yang telah diprogram sebelumnya dan tidak dapat meningkatkan diri sendiri. Contohnya adalah pemain catur Deep Blue milik IBM, yang dilatih untuk bermain catur pada tingkat ahli tetapi tidak mengingat permainan sebelumnya.
- Sistem AI dengan memori terbatas menyimpan informasi untuk jangka waktu singkat dan menggunakan informasi tersebut untuk meningkatkan kinerja mereka. Mereka belajar dari pengalaman masa lalu atau data historis untuk meningkatkan kemampuan pengambilan keputusan mereka. Namun, karena memori mereka terbatas, mereka biasanya membuang informasi ketika informasi tersebut tidak lagi relevan. Misalnya, kendaraan otonom menggunakan AI dengan memori terbatas untuk memproses data lalu lintas, kondisi jalan, dan pola mengemudi sebelumnya. Mereka menggunakan informasi ini untuk navigasi dan keputusan tentang kecepatan dan arah, tetapi mereka segera membuang informasi karena keterbatasan sumber daya.
- Sistem AI penalaran memiliki kemampuan canggih untuk menganalisis dan menghubungkan pola serta membuat prediksi yang tepat. Sebagian besar LLM (*Learning Learning Module*) yang umum saat ini termasuk dalam kategori ini. Mereka bahkan mampu lulus ujian standar (misalnya, ujian pengacara) dengan kinerja yang melebihi rata-rata manusia.
- Sistem AI teori pikiran adalah bentuk AI tingkat lanjut yang bersifat teoretis. AI ini dapat memahami dan mensimulasikan emosi, kepercayaan, niat, dan pikiran manusia. Di masa depan, sistem AI dapat berinteraksi dengan manusia dengan cara yang mirip dengan hubungan manusia, termasuk memahami dan menanggapi emosi. Contoh fiktifnya adalah komputer HAL-9000 dalam film 2001: *A Space Odyssey.*
- AI yang sadar diri adalah tingkat AI tertinggi yang memiliki kesadaran dan pemahaman diri. AI yang sadar diri akan mampu mengenali dirinya sendiri di lingkungannya dan memahami keberadaannya sendiri. Seperti AI supercerdas, AI yang sadar diri saat ini

masih merupakan konsep teoretis. Contoh AI yang sadar diri adalah robot humanoid yang mampu melakukan introspeksi, memahami keterbatasannya sendiri, dan bahkan mempertanyakan tujuannya.

- Robotika adalah jenis sistem AI yang mampu melakukan tugas-tugas kinetik, seperti membawa dan memindahkan objek, serta tugas-tugas yang lebih canggih, seperti memasak makanan.

Teknik AI

Cara lain untuk mengklasifikasikan sistem AI adalah berdasarkan teknik atau metode yang digunakan untuk melakukan tugas. Metode-metode ini membantu kita memahami cara kerja sistem AI. Ini termasuk:

- Pembelajaran mesin (ML) adalah teknik yang memungkinkan sistem untuk belajar dari data, meningkatkan kinerja, dan membuat keputusan tanpa pemrograman eksplisit. Dalam ML, algoritma dilatih pada kumpulan data besar dan "belajar" untuk membuat prediksi atau keputusan berdasarkan pola yang diidentifikasi dalam data. Filter spam, misalnya, menggunakan ML untuk menganalisis email masa lalu dan mengidentifikasi pola (seperti kata-kata tertentu atau perilaku pengirim) untuk membantu mengklasifikasikan email baru sebagai spam.
- Pembelajaran mendalam adalah sub-bidang pembelajaran mesin yang menggunakan jaringan saraf tiruan yang terdiri dari beberapa lapisan. Sistem pembelajaran mendalam dirancang untuk memproses volume data yang besar, seringkali tidak terstruktur, dan sangat efektif pada tugas-tugas seperti pengenalan suara dan gambar. Contoh yang baik adalah Google Translate, yang dapat menerjemahkan teks lisan atau tulisan ke berbagai bahasa.
- Sistem pemrosesan bahasa alami (NLP) dapat memahami, menafsirkan, dan menghasilkan bahasa manusia. NLP menggabungkan linguistik dan pembelajaran mesin untuk memproses teks dan ucapan, membuat interaksi antara manusia dan mesin lebih intuitif. Contohnya adalah ChatGPT, chatbot yang menggunakan NLP untuk memproses dan menanggapi masukan pengguna dalam bahasa manusia.
- Pembelajaran terawasi adalah jenis pelatihan sistem AI yang menggunakan data berlabel. Misalnya, filter spam berbasis ML dilatih dengan contoh pesan spam dan pesan yang sah. Dalam contoh lain, robot dilatih dengan kumpulan data gambar untuk kemudian mengidentifikasi berbagai jenis objek.
- Pembelajaran tak terawasi adalah jenis pelatihan sistem AI yang menggunakan data tak berlabel. Sistem AI diharapkan dapat mendeteksi pola sendiri. Misalnya, dengan diberikan basis data hubungan pelanggan, sistem AI diarahkan untuk mengklasifikasikannya sesuai dengan yang dianggapnya tepat.
- Pembelajaran semi-terawasi adalah jenis pelatihan sistem AI yang menggunakan sejumlah kecil data berlabel dan sejumlah besar data tak berlabel. Sistem AI pertama-tama belajar dari data berlabel dan kemudian dilatih pada data tak berlabel untuk memprediksi label. Seiring waktu, kinerja akan meningkat. Misalnya, sebuah organisasi sedang melatih sistem pengenalan gambar berbasis AI tetapi hanya memiliki sedikit

koleksi gambar untuk dilatih. Sistem AI dilatih pada data berlabel dan terus belajar dengan data yang tidak berlabel.

- Pembelajaran penguatan (*reinforcement learning*) adalah jenis pembelajaran mesin di mana agen AI belajar melalui interaksi dan menerima umpan balik berupa imbalan atau hukuman. Tujuannya adalah agar sistem AI memaksimalkan imbalannya melalui uji coba dan kesalahan. Contohnya adalah robot yang belajar menavigasi labirin.
- Sistem pakar (*expert systems*) adalah sistem AI yang mensimulasikan kemampuan pengambilan keputusan seorang pakar manusia di bidang tertentu, menggunakan aturan dan basis pengetahuan untuk membuat kesimpulan dan memecahkan masalah kompleks. Misalnya, sistem diagnostik medis dapat menganalisis gejala pasien dan riwayat medis untuk menyarankan diagnosis atau pengobatan.
- Pembelajaran tanpa contoh (*zero-shot learning*/ZSL) adalah kasus penggunaan pembelajaran mesin di mana model AI dilatih untuk mengenali dan mengkategorikan objek atau konsep tanpa pernah melihat contoh kategori atau konsep tersebut sebelumnya.

Memahami definisi-definisi ini dan definisi lainnya sangat penting untuk menavigasi dan mengatur AI. Karena beberapa profesional tata kelola AI tidak memiliki pengetahuan teknis yang mendalam, mereka harus memahami konsep-konsep seperti ini untuk membantu mereka mencapai kredibilitas yang lebih baik dengan rekan-rekan teknis mereka, sambil berupaya mencapai hasil bisnis yang lebih baik.

1.2 RISIKO DAN KERUGIAN YANG DITIMBULKAN OLEH AI

AI telah membawa perubahan transformatif di banyak industri dengan meningkatkan efisiensi, menciptakan peluang baru, dan memecahkan masalah yang kompleks. Namun, seiring AI semakin terintegrasi ke dalam bisnis dan masyarakat, AI juga menimbulkan risiko dan potensi kerugian. Memahami risiko-risiko ini sangat penting bagi para profesional yang mengelola dan mengatur sistem AI.

Mari kita jelajahi berbagai jenis risiko dan kerugian yang dapat ditimbulkan AI bagi individu, kelompok, organisasi, dan masyarakat. Setiap risiko dijelaskan dengan contoh untuk memberikan kejelasan dan konteks.

Risiko dan Kerugian bagi Individu

Potensi risiko dan kerugian bagi individu dapat berupa dua bentuk: ketika individu berinteraksi langsung dengan AI dan ketika organisasi menggunakan pemrosesan dan pengambilan keputusan berbasis AI dengan informasi tentang individu.

- **Pelanggaran Privasi**: Salah satu risiko paling signifikan yang ditimbulkan AI bagi individu adalah potensi pelanggaran privasi. Sistem AI, khususnya yang bergantung pada kumpulan data besar dan sistem *Retrieval Augmented Generation* (RAG), dapat mengekspos informasi pribadi yang sensitif, yang mengakibatkan pelanggaran privasi. AI yang telah dilatih atau memiliki akses ke informasi pribadi atau informasi identitas pribadi (PII) tanpa anonimisasi berpotensi membocorkan data sensitif ketika diminta dengan pertanyaan terkait. Misalnya, jika AI telah terpapar nama pengguna, kata sandi,

atau alamat email, ia mungkin dapat mereproduksi informasi ini dalam responsnya. Demikian pula, ia dapat mencoba memprediksi detail login pengguna ketika ditanya. Contoh lain adalah sistem pengenalan wajah yang digunakan oleh toko ritel untuk melacak pergerakan pelanggan. Sistem ini dapat mengumpulkan data biometrik tanpa persetujuan, yang menimbulkan kekhawatiran tentang pengawasan dan privasi pribadi.

- **Bias dan diskriminasi**: Sistem AI dapat melanggengkan atau bahkan memperkuat bias sosial yang ada, yang dapat merugikan individu, terutama mereka yang berasal dari kelompok marginal. Hal ini terjadi ketika model AI dirancang atau dilatih secara tidak tepat menggunakan data yang bias. Contoh terbaru termasuk sistem pengambilan keputusan berbasis AI yang membantu menyaring dan memenuhi syarat kandidat pekerjaan, pemohon pinjaman, dan penerima bantuan keuangan. Organisasi seringkali tidak menyadari bias dalam sistem mereka sampai para korban angkat bicara.
- **Risiko keamanan siber**: Banyak organisasi TI melakukan pekerjaan yang kurang memadai dalam melindungi sistem mereka dan data sensitif yang mereka simpan dan proses. Hal ini tampaknya lebih umum terjadi ketika organisasi memperkenalkan teknologi baru seperti AI mereka bersemangat untuk menggunakan mainan baru tanpa mempertimbangkan perlindungan yang memadai. Individu menjadi korban pencurian identitas, kerugian finansial, dan terungkapnya informasi pribadi mereka. Saat organisasi berlomba untuk membangun dan menerapkan sistem AI, praktik keamanan mendasar seringkali diabaikan. Kita melihat organisasi gagal besar; mulai dari melatih dan menyempurnakan AI pada data yang tidak dianonimkan hingga menerapkannya dalam sistem yang kurang memiliki kontrol keamanan yang tepat dan melanggar prinsip-prinsip seperti hak akses minimal dan pemisahan tugas. Pada akhirnya, individu lah yang menanggung akibatnya, karena privasi dan keamanan data mereka seringkali diabaikan.
- **Hilangnya visibilitas, otonomi, dan kendali**: Sistem pengambilan keputusan dan pemrosesan berbasis AI seringkali menghilangkan peran manusia, mengurangi otonomi dan kemampuan seseorang untuk memengaruhi hasil. Layanan kesehatan dan keuangan adalah dua industri di mana orang merasa diperlakukan seperti angka, bukan sebagai manusia. Misalnya, alat diagnostik berbasis AI dapat mengesampingkan keputusan penyedia layanan kesehatan atau salah menafsirkan kondisi pasien, menyebabkan individu kehilangan kendali atas rencana perawatan mereka dan berpotensi menyebabkan hasil kesehatan yang buruk. Situasi ini dapat terjadi lebih jauh di rantai pasokan ketika perusahaan asuransi kesehatan harus menyetujui suatu prosedur atau perawatan sebelum dapat dimulai.

Saya baru saja membahas sebagian kecilnya. Seiring dengan semakin meluasnya penggunaan AI di lebih banyak organisasi dan semakin banyak fungsinya, saya pikir kemungkinan besar kita akan melihat lebih banyak risiko seperti ini.

Risiko dan Kerugian bagi Kelompok

Selain dampak pada individu, AI juga dapat menimbulkan konsekuensi signifikan bagi kelompok, seperti komunitas, keluarga, dan jaringan sosial. AI menimbulkan risiko bagi kelompok yang cenderung lebih sistemik, dengan implikasi sosial yang luas.

- **Penguatan ketidaksetaraan sosial**: Sistem AI yang beroperasi dalam skala besar dapat memperburuk ketidaksetaraan yang ada. Sistem AI pengambilan keputusan yang memperoleh data profil dari media sosial dapat menunjukkan bias yang memengaruhi kelompok tertentu. Misalnya, sistem kualifikasi pinjaman berbasis AI mungkin menawarkan suku bunga yang lebih tinggi atau menolak pinjaman kepada individu di komunitas atau kelompok tertentu.
- **Erosi kohesi sosial**: Platform berbasis AI dapat secara tidak sengaja menciptakan ruang gema dengan sedikit keragaman ide, yang lebih mengutamakan konten sensasional yang meningkatkan polarisasi dan mengurangi toleransi terhadap sudut pandang. Misalnya, beberapa platform media sosial dapat memprioritaskan konten yang sarat emosi, menciptakan lebih banyak perpecahan dan mengikis kepercayaan masyarakat.
- **Risiko keamanan siber**: Adopsi AI lebih luas di infrastruktur penting (transportasi, utilitas, perawatan kesehatan) dan bisnis. Seperti yang telah disebutkan sebelumnya, organisasi tidak terlalu mahir dalam mengamankan teknologi baru seperti AI sampai masalah keamanan muncul. Misalnya, sebuah bank besar mungkin menerapkan chatbot berbasis AI yang dapat ditipu untuk mengungkapkan informasi tentang pelanggan lain. Risiko terkait keamanan siber dibahas di seluruh buku ini.

Seperti sebelumnya, ini adalah contoh-contoh penting tetapi bukan satu-satunya.

Risiko dan Kerugian bagi Organisasi

Organisasi mengadopsi sistem AI untuk meningkatkan efisiensi, kualitas, dan pengambilan keputusan serta untuk mengurangi biaya operasional tertentu. AI menimbulkan risiko unik bagi organisasi yang menggunakannya, termasuk:

- **Risiko operasional**: Organisasi yang mengandalkan AI untuk proses-proses penting, termasuk manajemen inventaris, deteksi penipuan, dan layanan pelanggan, rentan terhadap gangguan jika sistem AI gagal atau berperilaku tidak terduga. Semakin integral sistem AI, semakin besar potensi dampak kegagalan. Misalnya, kerusakan sistem inventaris toko ritel yang digerakkan oleh AI dapat mengakibatkan terlalu banyak atau terlalu sedikit stok beberapa produk, yang memengaruhi kepuasan pelanggan, pendapatan, dan kerusakan merek.
- **Risiko hukum dan peraturan**: Hukum dan peraturan mengenai privasi dan AI terus muncul dan berubah. Organisasi yang tidak memperhatikan perkembangan ini berisiko melanggar peraturan yang relevan, beberapa di antaranya dapat menyebabkan sanksi publik dan kerusakan reputasi. Misalnya, organisasi yang kurang memiliki tata kelola privasi data yang efektif dapat melanggar hukum privasi jika sistem yang ditingkatkan AI-nya memproses PII secara ilegal. Di sisi lain, badan pengatur di seluruh dunia juga kesulitan untuk mengikuti perkembangan dan penerapan sistem AI yang pesat.

- **Risiko keuangan**: Organisasi yang menggunakan AI dalam manajemen sumber daya atau sistem pengambilan keputusan mungkin mengalami hasil yang buruk. Lebih lanjut, mengadopsi teknologi baru seperti AI dapat melebihi perkiraan biaya, yang memengaruhi hasil keuangan. Misalnya, organisasi yang menggunakan perdagangan sekuritas berbasis AI dapat mengalami kerugian signifikan jika sistem AI salah menilai kondisi dan peristiwa pasar.
- **Risiko Keamanan Siber**: Organisasi semakin banyak mengadopsi alat keamanan siber berbasis AI untuk meningkatkan pertahanan mereka dan mengurangi kemungkinan atau dampak serangan siber. Seperti jenis sistem berbasis AI lainnya, sangat penting bagi organisasi untuk memahami bagaimana sistem berbasis AI seharusnya bekerja, dan melakukannya secara konsisten. Misalnya, sebuah organisasi membeli sistem pencegahan intrusi jaringan berbasis AI, yang memblokir lalu lintas jaringan dari mitra bisnis penting, mengakibatkan pemadaman selama berjam-jam dan publisitas negatif. Lebih lanjut, organisasi harus melindungi sistem AI mereka dari peracunan yang dapat memungkinkan mereka menghasilkan output yang salah.
- **Risiko Rantai Pasokan**: Organisasi yang menggunakan sistem AI yang dipasok vendor berisiko terhadap serangan terhadap vendor tersebut, di mana penyerang memasukkan kode berbahaya ke dalam perangkat lunak yang didistribusikan kepada pelanggannya. Serangan SolarWinds tahun 2020 adalah contoh klasik dari risiko ini.
- **Risiko Ketidaksesuaian**: Karena literasi AI yang terbatas dan kurangnya pemahaman tentang bagaimana model ini beroperasi, perilaku yang tampak tidak berbahaya di lingkungan pengujian dapat menyebabkan konsekuensi yang tidak diinginkan saat diterapkan. Misalnya, AI yang bertugas memaksimalkan produktivitas, yang diukur berdasarkan jumlah barang yang diproduksi per pekerja, mungkin menyimpulkan bahwa mengurangi jumlah pekerja adalah cara paling efektif untuk mencapai tujuan tersebut.
- **Risiko reputasi**: Jika terwujud, semua risiko yang disebutkan dalam bagian ini dapat menyebabkan pengungkapan publik, yang mengakibatkan reaksi negatif publik dan kerusakan merek. Misalnya, sistem iklan bertarget yang digerakkan oleh AI dapat memilih iklan yang "tidak sensitif", yang menyebabkan protes publik dan kerusakan reputasi.

Risiko dan kerugian bagi organisasi dalam bagian ini dapat dipertimbangkan dari perspektif organisasi, yang mencakup risiko yang sama bagi individu dan kelompok yang dijelaskan di bagian sebelumnya. Misalnya, sebuah organisasi yang menggunakan alat penyaringan lamaran kerja berbasis AI dapat menyebabkan kerugian bagi individu yang diperlakukan tidak adil, serta bagi organisasi itu sendiri, jika kerugian tersebut dipublikasikan.

Risiko dan Kerugian bagi Masyarakat

Dampak AI terhadap masyarakat adalah yang paling mendalam, karena risiko dan kerugian yang ditimbulkan oleh AI dapat membentuk kembali struktur sosial, ekonomi, dan politik. Dampak ini dapat melampaui lingkup organisasi dan kelompok individu, memengaruhi seluruh negara, wilayah, dan sistem global.

- ✓ **Perpindahan pekerjaan**: Penggunaan otomatisasi dalam organisasi dapat menyebabkan perpindahan pekerja di berbagai industri, termasuk manufaktur, transportasi, teknologi informasi, dan layanan pelanggan. Memang, saat saya menulis bab ini, saya melihat artikel berita yang meramalkan penurunan permintaan untuk insinyur perangkat lunak, transportasi, perwakilan layanan pelanggan, dan pekerjaan lainnya. Lebih lanjut, beberapa organisasi tidak lagi merekrut personel tingkat pemula di beberapa industri, yang menyebabkan lulusan perguruan tinggi baru kesulitan mencari pekerjaan. Misalnya, adopsi kendaraan otonom mengurangi permintaan akan pengemudi.
- ✓ **Ketidaksetaraan ekonomi**: Bias dalam data pelatihan dapat mengakibatkan perlakuan yang tidak setara terhadap individu, yang mengakibatkan atau memperburuk ketidaksetaraan ekonomi. Misalnya, pelamar pekerjaan dan pinjaman dapat diperlakukan secara tidak setara.
- ✓ **Erosi otonomi dan agensi manusia**: Masyarakat dapat menjadi terlalu bergantung pada pengambilan keputusan otomatis yang didukung AI, yang selanjutnya mengikis otonomi manusia, ketika keputusan tentang individu, kelompok, dan masyarakat dapat dibuat oleh algoritma daripada manusia. Misalnya, pemerintah dapat semakin bergantung pada algoritma AI untuk membuat keputusan kebijakan publik, daripada melibatkan pemilih dengan cara tradisional.
- ✓ **Tantangan etika dan moral**: Seiring sistem AI menjadi lebih kuat dan diandalkan untuk membuat keputusan, dilema etika akan muncul, seperti ketika sistem ditugaskan untuk membuat keputusan hidup atau mati. Tantangan ini menjadi sangat menonjol ketika keputusan AI bertentangan dengan nilai atau norma masyarakat. Misalnya, taksi otonom dengan penumpang dapat menghadapi kecelakaan yang tak terhindarkan di mana penumpangnya atau pejalan kaki di dekatnya dapat terluka atau tewas. Bagaimana seharusnya kendaraan otonom tersebut merespons?
- ✓ **Keamanan siber dan privasi**: Ruang lingkup potensial masalah keamanan siber dan privasi hanya dibatasi oleh ruang lingkup sistem itu sendiri. Jika terjadi kerusakan terkait AI, utilitas publik skala besar, sistem kontrol, dan sistem pengawasan dapat menimbulkan kerugian yang cukup besar, termasuk cedera tubuh dan kematian. Misalnya, jaringan listrik besar yang menggunakan pengambilan keputusan AI untuk mengelola beban listrik dapat salah menafsirkan sinyal yang masuk, yang mengakibatkan pemadaman listrik yang meluas.
- ✓ **Manipulasi opini publik**: Platform media sosial saat ini menggunakan algoritma berbasis AI untuk mengontrol umpan informasi guna meningkatkan keterlibatan dan pendapatan iklan. Potensi kekuatan sosial dan politik mungkin terlalu kuat untuk ditolak. Misalnya, bertentangan dengan hukum, pemerintah nasional dapat memanipulasi platform media sosial populer untuk menguntungkan kandidat pemilihan tertentu.

Bagian ini meninjau potensi dampak AI pada tingkat individu, kelompok, organisasi, dan masyarakat. Risiko yang dibahas di sini jarang hanya terbatas pada satu tingkat, tetapi akan

meluas dan berdampak pada banyak area lainnya. Misalnya, perusahaan besar yang menggunakan fungsi penyaringan karyawan berbasis AI untuk mengidentifikasi kandidat yang cocok untuk wawancara mungkin telah melatih algoritma seleksinya secara tidak tepat, yang mengakibatkan kerugian bagi pelamar tertentu, kelompok orang, organisasi itu sendiri, dan berpotensi bagi masyarakat (setidaknya di wilayah tempat organisasi tersebut beroperasi).

Pahami saya: Saya tidak menentang AI. Sebaliknya, saya seorang optimis dan percaya bahwa AI menghadirkan peluang luar biasa bagi individu, kelompok, organisasi, dan masyarakat. Tetapi di mana pun Anda menemukan peluang besar, Anda juga menemukan risiko yang signifikan. Saya percaya bahwa keseimbangan kosmik ini tidak dapat dihindari.

Bagi pembaca yang (atau akan) bertanggung jawab atas tata kelola AI, saya harap bagian ini telah memberikan informasi praktis tingkat tinggi untuk membantu Anda lebih memahami berbagai jenis risiko yang terkait dengan AI. Bab-bab selanjutnya dalam buku ini akan membahas risiko-risiko ini dan risiko lainnya secara lebih rinci.

Kita Pernah Melihat Kegilaan Ini Sebelumnya

Dikatakan bahwa tidak ada yang baru di bawah matahari. Ketika berbicara tentang AI, dengan semua gembar-gembor, kegilaan, dan janji untuk menyelesaikan setiap masalah, ini bukanlah pertama kalinya teknologi baru membuat orang-orang yang biasanya rasional menjadi sedikit gila. Menengok ke belakang, kita memiliki munculnya komputer pribadi pada tahun 1980-an, Internet pada tahun 1990-an, komputasi awan pada tahun 2000-an, dan sekarang AI. Setiap kasus melibatkan organisasi yang bergegas untuk menerapkan teknologi ini, dengan sedikit memperhatikan pemikiran dan pengambilan keputusan yang rasional. Organisasi dapat mabuk dengan prospek keunggulan sebagai pelopor. Kejatuhan dot-com di awal tahun 2000-an adalah bukti bahwa organisasi bergegas untuk "online" tanpa terlebih dahulu membuat rencana bisnis yang matang dan mendefinisikan masalah spesifik yang harus dipecahkan. Dengan AI, sama tetapi juga berbeda, AI tertanam dalam segala hal saat ini, dan menambahkan AI tanpa memahami risikonya dengan cermat penuh dengan bahaya yang lebih besar daripada iterasi teknologi baru yang canggih sebelumnya.

1.3 KARAKTERISTIK AI YANG MEMBUTUHKAN TATA KELOLA

Sistem AI telah berkembang pesat, terintegrasi ke dalam setiap tingkatan kehidupan manusia, dari kehidupan individu sehari-hari hingga organisasi dan pemerintah nasional. Seiring AI terus mentransformasi pemerintah dan industri, kompleksitasnya dan potensi konsekuensi dari penerapannya memerlukan pembentukan struktur tata kelola untuk mengelolanya dan memastikan hasil yang menguntungkan. Tidak seperti sistem informasi tradisional, AI memperkenalkan tantangan baru yang membutuhkan strategi tata kelola yang disesuaikan untuk memastikan penggunaan yang etis, meminimalkan risiko, dan mengoptimalkan manfaat. Dengan kekuatan besar datang tanggung jawab besar. Tata kelola menyediakan pengaman.

Bagian ini mengeksplorasi karakteristik unik AI yang memerlukan pendekatan komprehensif terhadap tata kelola. Karakteristik ini meliputi:

- Kompleksitas
- Ketidaktransparan
- Otonomi
- Kecepatan dan skala
- Potensi bahaya dan penyalahgunaan
- Ketergantungan data
- Privasi
- Hak kekayaan intelektual
- Output probabilistik vs. deterministik

Masing-masing dibahas lebih detail di bagian ini.

Kompleksitas

Sistem AI seringkali sangat kompleks terkait model dasarnya dan lingkungan kompleks tempat mereka beroperasi. Kompleksitas ini membutuhkan kerangka kerja tata kelola yang dapat mengatasi kerumitan teknis dan operasionalnya. Kompleksitas muncul dari teknik-teknik canggih seperti jaringan saraf dan pembelajaran mendalam, di mana hubungan antar input dan output tidak mudah dipahami.

Sebagai contoh, industri perawatan kesehatan menggunakan alat diagnostik bertenaga AI dengan algoritma pembelajaran mendalam yang dilatih pada sejumlah besar data medis. Sistem ini dapat mengidentifikasi pola yang mungkin terlewatkan oleh manusia, dan kompleksitasnya menyulitkan penyedia layanan kesehatan untuk memahami secara tepat mengapa sistem tersebut membuat diagnosis tertentu. Hal ini menimbulkan kekhawatiran tentang kemampuan menjelaskan, kemampuan menafsirkan, akuntabilitas, dan transparansi.

Struktur tata kelola harus memastikan bahwa sistem AI dapat dijelaskan, bahwa manajemen bertanggung jawab atas semua hasil, dan bahwa sistem dapat diaudit. Ini termasuk mengembangkan standar transparansi dan praktik dokumentasi, memastikan bahwa sistem AI dapat dipantau secara efektif untuk menilai dan mengurangi risiko.

Opasitas

Beberapa sistem AI, pada dasarnya, adalah "kotak hitam," terutama yang menggunakan model pembelajaran mesin dan pembelajaran mendalam. Pada intinya, sulit untuk memahami bagaimana input diubah menjadi output, sehingga proses pengambilan keputusan sistem AI menjadi buram dan misterius. Opasitas ini menghadirkan tantangan tata kelola yang signifikan, terutama di bidang-bidang berisiko tinggi seperti keuangan, perawatan kesehatan, dan peradilan pidana, di mana keputusan yang didasarkan pada prediksi AI dapat memiliki konsekuensi serius. Kebalikan dari opasitas adalah kemampuan menjelaskan, di mana manusia dapat memahami dan menjelaskan bagaimana sistem AI membuat keputusan tertentu atau menciptakan output tertentu. Kemampuan menjelaskan mirip dengan kemampuan menafsirkan, yaitu kemampuan sistem AI untuk memahami hubungan antara pola dalam data. Misalnya, sistem persetujuan pinjaman berbasis AI bank mungkin menolak permohonan tanpa menjelaskan secara jelas faktor-faktor yang memengaruhi keputusannya. Jika penolakan tersebut didasarkan pada data yang bias, opasitas sistem AI menyulitkan pemohon untuk membantah keputusan tersebut atau memahami mengapa keputusan itu

dibuat. Demikian pula, bank mungkin kesulitan menjelaskan secara tepat mengapa pinjaman ditolak, yang berpotensi membahayakan bank karena masalah regulasi.

Struktur tata kelola harus memastikan bahwa sistem AI tertentu menggabungkan penjelasan dan transparansi dalam desain dan pelatihannya. Ini termasuk mensyaratkan bahwa sistem AI memberikan alasan yang jelas dan mudah dipahami untuk keputusan, terutama di bidang seperti keuangan dan perawatan kesehatan, di mana transparansi dapat melindungi hak individu, mencegah diskriminasi, dan melindungi organisasi dari tindakan hukum.

Otonomi

Beberapa sistem AI bersifat otonom, artinya mereka membuat keputusan atau mengambil tindakan tanpa inisiasi atau intervensi manusia, meningkatkan kompleksitas pengelolaan sistem AI, karena sulit untuk memprediksi bagaimana agen AI akan berperilaku dalam setiap situasi. Sistem otonom mungkin dapat berevolusi dan beradaptasi dari waktu ke waktu, berdasarkan data baru, menambah tantangan pada tata kelola. Misalnya, kendaraan otonom dapat membuat keputusan sepersekian detik untuk menghindari tabrakan dengan berbelok, yang berpotensi membahayakan pejalan kaki. Proses pengambilan keputusan yang digerakkan AI pada kendaraan bersifat otonom dan tidak dapat dengan mudah ditimpa atau dipengaruhi oleh operator manusia (jika ada) selama momen-momen kritis.

Struktur tata kelola harus mengidentifikasi batasan yang tepat terhadap tindakan sistem AI, memastikan langkah-langkah keamanan diterapkan, dan memantau kinerja secara berkala. Fungsi tata kelola harus memberlakukan kebijakan yang jelas untuk mendefinisikan apa yang dapat dan tidak dapat dilakukan oleh sistem AI, serta menangani dan menetapkan akuntabilitas atas keputusan yang dibuat oleh sistem otonom.

Kecepatan dan Skala

Sistem AI memproses data dan membuat keputusan jauh lebih cepat daripada manusia, memungkinkan mereka untuk beroperasi dalam skala besar di berbagai aplikasi. Kecepatan operasi sistem AI merupakan kekuatan sekaligus risiko. Kemampuan AI untuk memproses sejumlah besar data dengan kecepatan tinggi memungkinkan untuk memecahkan masalah yang mustahil untuk diatasi manusia secara manual. Namun, kecepatan dan skala AI juga dapat menyebabkan konsekuensi yang tidak diinginkan (terutama biaya listrik dan komputasi) jika sistem tidak diatur dengan benar. Misalnya, algoritma AI platform media sosial secara otomatis merekomendasikan konten kepada jutaan pengguna, mengoptimalkan keterlibatan dan, pada akhirnya, pendapatan iklan. Namun, ini juga dapat menyebarkan informasi yang salah dan berbahaya sebelum terdeteksi dan ditangani.

Struktur tata kelola harus dipersiapkan untuk kecepatan dan skala serta dampak sumber daya dari pengambilan keputusan AI, yang memerlukan pemantauan waktu nyata, protokol respons, dan kontrol yang tepat untuk mencegah dan mengurangi risiko skala besar seperti informasi yang salah, penipuan keuangan, dan pelanggaran keamanan.

Dampak Global AI

Menurut Scientific American, pusat data (termasuk yang menjalankan sistem AI) menyumbang sekitar 1,5% dari penggunaan listrik global, dan ini diproyeksikan akan berlipat ganda pada tahun 2030. Pada saat itu, AI akan mengonsumsi listrik sebanyak seluruh negara Jepang. Di Amerika Serikat, pusat data yang digerakkan oleh AI diperkirakan akan menyumbang setengah dari peningkatan penggunaan listrik pada tahun 2030.

Potensi Bahaya dan Penyalahgunaan

Seperti yang telah dibahas sebelumnya dalam bab ini, sistem AI berpotensi digunakan untuk tujuan yang berbahaya dan menyebabkan konsekuensi negatif yang tidak diinginkan. Baik melalui pelaku jahat atau perilaku yang tidak terduga, potensi bahaya AI membutuhkan struktur tata kelola yang kuat untuk mendeteksi dan mencegah penyalahgunaan serta mengelola risiko. Sistem AI dapat diuji pada fase desain untuk mendeteksi potensi bahaya. Misalnya, sistem pengawasan berbasis AI dapat digunakan oleh pemerintah otoriter untuk memantau aktivitas warga negara, menekan perbedaan pendapat, dan melanggar hak privasi (ini bukan teori tetapi sudah terjadi di beberapa negara). Demikian pula, senjata bertenaga AI dalam aplikasi militer dapat mengalami malfungsi atau disalahgunakan, yang menyebabkan masalah etika dan keselamatan.

Struktur tata kelola harus menetapkan perlindungan untuk mendeteksi dan mencegah penggunaan AI yang berbahaya, termasuk mekanisme pengawasan etika, akuntabilitas, dan kontrol keamanan. Di beberapa yurisdiksi, seperti Uni Eropa, kerangka peraturan mendefinisikan penggunaan AI yang diizinkan dan memastikan bahwa AI tidak digunakan dengan cara yang melanggar hak asasi manusia atau menimbulkan risiko terhadap keselamatan publik.

Ketergantungan Data

Sebagian besar jenis sistem AI sangat bergantung pada data, terkadang dalam jumlah yang sangat besar. Kualitas, volume, dan keragaman data pelatihan AI secara langsung memengaruhi kinerja dan keandalan sistem ini. Data yang buruk atau bias dapat menyebabkan hasil yang tidak akurat atau tidak adil. Misalnya, sistem AI yang digunakan dalam peradilan pidana, seperti alat penjatuhan hukuman, dapat dilatih dengan data historis yang bias, yang menyebabkan hasil yang diskriminatif. Dalam contoh lain, jika data penangkapan historis secara tidak proporsional mewakili kelompok demografis tertentu, sistem AI dapat secara tidak adil menandai individu dari kelompok tersebut sebagai memiliki risiko pengulangan tindak pidana yang lebih tinggi.

Struktur tata kelola harus menangani masalah kualitas dan keamanan data. Organisasi harus menetapkan standar untuk mengumpulkan, menyimpan, dan memproses data untuk memastikan data tersebut akurat, lengkap, dan terlindungi.

Privasi

Banyak sistem AI sangat bergantung pada data tentang karyawan, pelanggan, dan konstituen lainnya. Seringkali, kumpulan data ini berisi satu atau lebih elemen PII (Informasi Identitas Pribadi). Misalnya, organisasi mengumpulkan PII dari pelanggan dan membeli lebih

banyak dari broker data. Mereka dapat melatih sistem AI mereka dengan data ini untuk berbagai tujuan, termasuk perkiraan penjualan, pemasaran yang ditargetkan, preferensi produk dan layanan, dan analisis demografis. Beberapa organisasi mungkin melanggar kebijakan privasi mereka sendiri, pemberitahuan praktik privasi (NOPP), dan/atau hukum privasi yang berlaku jika mereka tidak terlebih dahulu memeriksanya untuk menentukan apa yang diizinkan.

Kita masih kekurangan metode yang andal untuk menghentikan pelatihan AI setelah AI tersebut menyerap data. Ini menimbulkan tantangan serius: setelah fase pelatihan selesai, mematuhi peraturan perlindungan data menjadi hampir mustahil. Ini menggarisbawahi poin penting: Data harus dianonimkan dan dilindungi sebelum dimasukkan ke dalam model AI.

Struktur tata kelola harus menangani masalah bias, keadilan, privasi, legalitas, dan perlindungan PII. Sangat penting untuk menetapkan standar pengumpulan, penyimpanan, dan pemrosesan data untuk menghindari bias, memastikan kepatuhan terhadap peraturan privasi (misalnya, GDPR), menyertakan pemberitahuan praktik privasi (NOPP), dan melindungi informasi sensitif.

Hak Kekayaan Intelektual

Beberapa sistem AI publik, termasuk LLM seperti ChatGPT, Copilot, Grok, dan Claude, diketahui melatih model AI mereka dengan sejumlah besar informasi, termasuk konten berhak cipta seperti artikel surat kabar, buku, gambar, dan karya seni. Hingga pertengahan tahun 2025, belum ada preseden hukum yang mapan di Amerika Serikat untuk menentukan apakah melatih sistem AI dengan konten berhak cipta itu etis atau legal. Sebagai contoh, Getty Images mengajukan gugatan terhadap Stability AI pada tahun 2023, dengan tuduhan bahwa Stability AI menggunakan jutaan foto Getty Images, beserta keterangan dan metadata, untuk melatih alat pembuatan gambar AI-nya tanpa izin.

Kerangka kerja hak cipta yang ada juga gagal melindungi gaya, teknik, dan metode pribadi seniman. Secara desain, AI saat ini mempelajari pola dalam data yang digunakan untuk melatihnya, yang berarti jika AI dilatih pada karya seniman tertentu, AI tersebut mungkin mengadopsi gaya mereka dan tidak selalu mereplikasi karya mereka secara persis. Dengan demikian, hukum hak cipta membuat seniman sangat tidak terlindungi.

Struktur tata kelola harus menerapkan langkah-langkah untuk memastikan bahwa semua data yang digunakan untuk sistem AI diperiksa hak ciptanya dan masalah kekayaan intelektual lainnya. Demikian pula, tata kelola AI juga perlu memastikan bahwa tenaga kerja suatu organisasi sangat menyadari aturan bisnis mengenai penggunaan model AI publik dan apakah data organisasi apa pun dapat diunggah ke dalamnya.

Output Probabilistik vs. Deterministik

Beberapa sistem AI, terutama jaringan saraf dan yang berbasis pembelajaran mesin, sering menghasilkan output probabilistik, artinya prediksi atau keputusan mereka didasarkan pada probabilitas daripada kepastian. Sebaliknya, output sistem deterministik tradisional sepenuhnya dapat diprediksi dengan input yang sama. Sifat probabilistik AI memperkenalkan unsur ketidakpastian, yang dapat menjadi tantangan ketika hasil memiliki konsekuensi yang signifikan dan harus dijelaskan. Misalnya, berdasarkan probabilitas, alat diagnostik AI

perawatan kesehatan memprediksi kemungkinan pasien mengembangkan kondisi tertentu. Prediksi tersebut tidak definitif, dan mengandalkannya dapat menyebabkan perawatan yang tidak perlu atau kesalahan diagnosis jika probabilitas disalahartikan atau akurasi sistem AI tidak cukup tinggi.

Struktur tata kelola harus mengatasi ketidakpastian yang melekat pada sistem AI probabilistik. Manajemen mungkin perlu mendefinisikan ambang batas yang dapat diterima untuk pengambilan keputusan AI, memberikan pedoman ketat untuk menafsirkan probabilitas, dan memastikan transparansi dalam pengambilan keputusan. Seperti selera risiko dan toleransi risiko, organisasi juga harus menentukan tingkat kepastian yang dibutuhkan dalam konteks risiko tinggi, sebagaimana didefinisikan oleh model klasifikasi AI mereka.

1.4 PRINSIP-PRINSIP AI YANG BERTANGGUNG JAWAB

Seiring AI semakin terintegrasi ke dalam kehidupan sehari-hari, masyarakat, dan sektor bisnis penting seperti energi, perawatan kesehatan, keuangan, dan peradilan pidana, prinsip-prinsip AI yang bertanggung jawab telah muncul sebagai prinsip inti untuk penggunaan AI yang etis dan efektif. AI yang bertanggung jawab memastikan bahwa sistem AI dirancang, dikembangkan, dan diterapkan untuk menghormati hak asasi manusia, mempromosikan keadilan, dan meminimalkan risiko sambil memberikan manfaat bagi individu, organisasi, dan masyarakat.

Bagian selanjutnya dari bagian ini mengidentifikasi dan menjelaskan prinsip-prinsip utama AI yang bertanggung jawab, termasuk etika, keadilan, keselamatan dan keandalan, privasi, keamanan, transparansi dan penjelasan, akuntabilitas, dan berpusat pada manusia. Prinsip-prinsip ini adalah elemen inti yang seharusnya mengatur akuisisi, pengembangan, dan penggunaan AI secara bertanggung jawab. Setiap prinsip diilustrasikan dengan contoh untuk menunjukkan penerapannya secara praktis.

Etika

Etika harus dianggap sebagai fondasi utama AI yang bertanggung jawab. Pertimbangan etika harus memandu pengembangan, penerapan, dan penggunaan sistem AI agar selaras dengan nilai-nilai inti manusia. Sistem AI harus dirancang untuk beroperasi dalam batasan etika yang mencerminkan norma-norma masyarakat dan prinsip-prinsip moral yang sehat.

Tantangan Etika

AI mengganggu norma etika dan sosial. Algoritma AI dapat membuat keputusan yang bertentangan dengan nilai-nilai moral, seperti memprioritaskan keuntungan di atas kesejahteraan manusia atau mengabaikan hak asasi manusia.

Seperti statistik, etika dapat diputarbalikkan untuk menyesuaikan hasil apa pun. Organisasi harus memastikan bahwa standar etika mereka didasarkan pada fondasi yang kokoh dan tidak didorong oleh agenda yang tidak berpusat pada manusia. Misalnya, organisasi militer menggunakan drone bertenaga AI yang membuat keputusan di medan perang, dengan implikasi etika yang menimbulkan kekhawatiran serius tentang akuntabilitas dan martabat manusia.

Penerapan Etika

AI yang etis membutuhkan pedoman yang komprehensif dan tepat untuk memastikan bahwa sistem AI mempromosikan kesejahteraan manusia dan tidak membahayakan atau mengeksploitasi pihak yang rentan. Sistem AI harus selaras dengan prinsip-prinsip etika yang tepat, termasuk hak asasi manusia, martabat, dan kebaikan sosial. Organisasi harus menugaskan audit etika, yang dilakukan oleh pihak eksternal yang objektif tanpa kepentingan dalam hasilnya. Penilaian dampak dan keterlibatan pemangku kepentingan sangat penting untuk mengidentifikasi dan mengatasi risiko etika di seluruh siklus pengembangan AI.

Pertanyaan untuk Pertimbangan Lebih Lanjut

Dapatkah sistem AI yang dikembangkan tanpa pertimbangan etis dianggap bertanggung jawab? Mengapa atau mengapa tidak?

Siapa, di dalam atau di luar organisasi, yang harus menetapkan norma-norma etika? Bagaimana norma-norma tersebut dapat ditentukan sebagai norma yang sahih dan dapat dipertahankan?

Bagaimana norma-norma etika berbeda di antara masyarakat? Dapatkah sistem AI dengan audiens global memenuhi norma-norma etika di semua lokasi?

Keadilan

Manusia tampaknya memiliki naluri bawaan untuk keadilan, seperti yang ditunjukkan oleh balita yang dengan cepat menunjukkan perbedaan dalam cara mereka diperlakukan dengan seruan, “Itu tidak adil!” Keadilan adalah komponen inti dari tatanan sosial kita dan sangat penting untuk AI yang bertanggung jawab.

Keadilan adalah prinsip inti dari AI yang bertanggung jawab yang mendorong kita untuk merancang dan menggunakan sistem AI yang tidak mendiskriminasi individu atau kelompok berdasarkan karakteristik seperti ras, jenis kelamin, afiliasi agama dan politik, etnis, status sosial ekonomi, dan disabilitas. Ketika tercapai, sistem AI meminimalkan atau menghilangkan bias dan mempromosikan hasil yang adil bagi semua.

Tantangan dengan Keadilan

Sistem AI dapat secara tidak sengaja melanggengkan atau bahkan memperkuat bias yang ada, terutama jika data pelatihannya mencerminkan ketidakadilan historis. Disengaja atau tidak, bias dapat menyebabkan hasil yang diskriminatif, terutama ketika AI digunakan dalam proses pengambilan keputusan, termasuk perekrutan, pemberian pinjaman, atau peradilan pidana. Misalnya, sebuah organisasi ritel besar menggunakan sistem bertenaga AI untuk menyaring lamaran pekerjaan baru. Setelah beberapa waktu, manajemen menemukan bahwa sistem penyaringan menolak pelamar perempuan jauh lebih banyak daripada pelamar laki-laki. Investigasi terhadap masalah ini menemukan bahwa data pelatihan yang digunakan dalam sistem tersebut terdiri dari resume dan profil karyawan yang sudah ada, yang sebagian besar dan secara tidak proporsional adalah laki-laki. Tidak semua masyarakat dan budaya memiliki norma yang sama mengenai keadilan. Organisasi multinasional mungkin perlu

memahami potensi perbedaan dan merancang proses dan sistem AI mereka sesuai dengan hal tersebut.

Penerapan Keadilan

Untuk memastikan bahwa sistem AI menunjukkan keadilan, organisasi harus memastikan bahwa data yang digunakan untuk melatih sistem AI beragam dan representatif, secara teratur mengaudit sistem AI untuk bias, dan menerapkan teknik seperti mitigasi bias untuk mencegah diskriminasi lebih lanjut. Sistem AI harus memastikan akses dan peluang yang adil bagi semua orang, terlepas dari karakteristik demografis mereka.

Keamanan dan Keandalan

Keamanan dan keandalan adalah prinsip-prinsip penting yang memastikan sistem AI aman bagi individu, kelompok, organisasi, dan masyarakat. Sistem AI harus berfungsi sebagaimana mestinya, bahkan ketika menghadapi masukan yang tidak terduga di lingkungan yang tidak dapat diprediksi. Sistem AI juga harus tangguh untuk menangani situasi yang terkait dengan kesalahan atau malfungsi.

Tantangan dengan Keamanan

Sistem AI, terutama sistem otonom yang mempertaruhkan nyawa, dapat menimbulkan risiko signifikan jika mengalami malfungsi atau membuat keputusan yang buruk. Taruhannya tinggi di area yang kritis terhadap keselamatan seperti perawatan kesehatan, tanggap darurat, dan transportasi otonom.

Pada kendaraan otonom, AI membuat keputusan navigasi dan mengemudi secara real-time. Kerusakan atau kegagalan pada sistem AI, seperti ketidakmampuan untuk mengidentifikasi pejalan kaki dalam kondisi pencahayaan yang buruk, dapat menyebabkan kecelakaan, cedera, dan bahkan kematian.

Aplikasi Keselamatan dan Keandalan

Sistem AI yang terkait dengan keselamatan jiwa harus dikembangkan dengan persyaratan yang ketat dan komprehensif. Sistem AI tersebut harus menjalani pengujian dan validasi yang ketat dalam berbagai kondisi. Lebih lanjut, sistem AI tersebut harus dirancang dengan pengaman kegagalan, kemampuan untuk diinterupsi, dan toleransi kesalahan. Organisasi harus melakukan pemantauan terus-menerus terhadap sistem AI dalam konteks keselamatan jiwa untuk mengidentifikasi dan mengatasi risiko yang muncul.

Privasi

Banyak sistem AI memproses sejumlah besar data pribadi sensitif dan PII (Informasi Identitas Pribadi). Untuk membangun dan mempertahankan kepercayaan, organisasi harus mengambil tindakan pencegahan ekstra agar semua penggunaan AI mematuhi peraturan privasi yang berlaku. Untuk memastikan privasi PII, organisasi harus dengan cermat melacak semua penggunaan PII dan menyediakan sarana untuk menanggapi pertanyaan, koreksi, dan permintaan penghapusan sesuai kebutuhan.

Tantangan Terkait Privasi

Bahkan sebelum AI, organisasi menghadapi tantangan untuk membatasi penggunaan data karyawan dan pelanggan untuk tujuan di luar apa yang diatur dalam kebijakan privasi dan peraturan yang berlaku. Beberapa organisasi bertindak seolah-olah tidak ada aturan mengenai

sistem AI dan data yang digunakan untuk melatihnya. Tanpa tata kelola data dan fungsi tata kelola AI yang kuat, perilaku yang salah dapat terus berlanjut tanpa terkendali sampai praktik-praktik ini dapat dikendalikan. Misalnya, aplikasi kesehatan seluler suatu organisasi yang menggunakan AI untuk melacak tujuan kebugaran dapat secara tidak sengaja mengekspos data kesehatan sensitif pengguna jika desainnya tidak dianalisis untuk kepatuhan terhadap kebijakan dan peraturan. Pelanggaran tersebut dapat menyebabkan sanksi publik dan kerusakan merek.

Penerapan Privasi

Sistem AI yang menggunakan PII (Informasi Identitas Pribadi) harus dirancang untuk mematuhi semua hukum privasi yang berlaku. Teknik privasi, termasuk minimalisasi data (yaitu, hanya mengumpulkan data yang diperlukan untuk tugas yang ada) dan deidentifikasi (berbagai teknik untuk membersihkan kumpulan data), harus diterapkan. Langkah-langkah keamanan, yang akan dibahas selanjutnya, harus menjadi bagian dari desain sistem AI.

Keamanan

Banyak sistem AI memproses sejumlah besar data sensitif, termasuk informasi pribadi, kekayaan intelektual, dan konten berhak cipta. Melindungi data ini dari akses dan kompromi yang tidak sah merupakan landasan kepercayaan dan sangat penting untuk hukum privasi yang berlaku (yang mensyaratkan penanganan dan penggunaan data pribadi yang tepat serta perlindungannya dari pencurian).

Melindungi informasi pribadi, kekayaan intelektual, dan konten lainnya merupakan hal mendasar melalui praktik keamanan siber yang telah lama ada. Buku ini merangkum teknik-teknik ini, yang dieksplorasi secara rinci dalam buku-buku lain yang diterbitkan oleh penulis. Sebagai contoh, sebuah organisasi ritel berupaya menerapkan AI untuk meningkatkan efisiensi fungsi layanan pelanggannya. Seperti banyak rekan sejawatnya, organisasi tersebut sudah kesulitan memahami di mana semua data pentingnya berada dan bagaimana data tersebut digunakan. Masalah ini semakin diperparah ketika sebagian data ini juga digunakan untuk melatih sistem AI-nya, yang menambah proliferasi data sensitif yang sulit dilindungi.

Tantangan Keamanan

Banyak organisasi kesulitan melindungi informasi sensitif dan sistem penting, karena keamanan siber seringkali diprioritaskan lebih rendah daripada profitabilitas, waktu pemasaran, dan keharusan kelangsungan bisnis lainnya. Organisasi yang kurang memiliki tata kelola TI dan keamanan siber yang efektif juga akan kesulitan dengan AI jika kemampuan dasar ini di bawah standar. Namun, organisasi menghadapi tekanan yang sangat besar untuk berinovasi dan mendorong efisiensi menggunakan AI, seperti yang mungkin sudah dilakukan oleh rekan-rekan mereka. Hal ini diperparah oleh fakta bahwa AI modern merusak batasan keamanan dan kepercayaan tradisional.

Aplikasi Keamanan

Organisasi harus terlebih dahulu menginventarisasi sistem informasi dan kumpulan data sensitif mereka, kemudian melakukan penilaian risiko untuk mengidentifikasi risiko potensi kompromi dan kehilangan. Kemudian, dengan dukungan manajemen, pengamanan tambahan dapat diterapkan atau ditingkatkan untuk membawa keamanan siber ke tingkat

kehati-hatian yang semestinya. Proses ini merupakan cetak biru standar, meskipun singkat, untuk keamanan siber modern.

Transparansi dan Kemampuan Penjelasan

Istilah transparansi secara luas merujuk pada kemampuan suatu organisasi untuk memahami bagaimana sistem AI-nya berfungsi. Lebih spesifiknya, interpretasi adalah kemampuan suatu organisasi untuk memahami bagaimana mereka mencapai kesimpulannya. Kemampuan penjelasan adalah karakteristik sistem AI yang memungkinkan suatu organisasi untuk memberikan justifikasi yang dapat dipahami atas tindakan yang dilakukan sistem tersebut. Prinsip-prinsip penting ini membantu membangun kepercayaan dan memastikan akuntabilitas, khususnya untuk sistem AI yang membuat keputusan yang berdampak.

Tantangan dengan Transparansi dan Kemampuan Penjelasan

Banyak model AI, termasuk jaringan saraf dan yang menggunakan pembelajaran mendalam, beroperasi sebagai "kotak hitam," sehingga sulit bagi suatu organisasi untuk memahami output dan keputusannya. Ketidakjelasan ini menyebabkan masalah kepercayaan dan tantangan terhadap keputusan sistem AI. Misalnya, sebuah bank komunitas baru-baru ini menerapkan alat AI untuk memeriksa permohonan pinjaman. Ketika sistem AI menolak permohonan pinjaman, organisasi tersebut mungkin tidak memberikan alasan yang ringkas untuk penolakan tersebut. Pemohon tidak puas dengan kurangnya penjelasan yang kredibel dan dapat mengeluh kepada regulator.

Penerapan Transparansi dan Keterjelasan

Untuk menghindari situasi di mana suatu organisasi tidak dapat menjelaskan secara memadai alasan di balik keputusan dan keluaran sistem AI, organisasi harus memprioritaskan pengembangan AI yang dapat dijelaskan (XAI), di mana model AI memberikan alasan yang dapat dipahami untuk keluarannya. Meskipun hal ini mungkin tampak menghambat organisasi yang ingin menyediakan kemampuan AI yang paling ampuh, penggunaan XAI dalam konteks pengambilan keputusan dapat membantu organisasi menghindari situasi yang dapat menyebabkan kerugian finansial, sosial, operasional, dan reputasi.

Akuntabilitas

Akuntabilitas adalah ciri organisasi yang memastikan bahwa individu dan organisasi bertanggung jawab atas hasil sistem AI mereka. Peran dan tanggung jawab untuk pengawasan dan penyelesaian masalah sistem AI sangat penting ketika masalah muncul.

Tantangan dengan Akuntabilitas

Sistem AI sering beroperasi secara otonom dan membuat keputusan secara independen, tetapi ini tidak membebaskan siapa pun dari tanggung jawab ketika terjadi kesalahan. Misalnya, sistem diagnostik yang ditingkatkan AI dari organisasi perawatan kesehatan membuat diagnosis yang salah, yang menyebabkan kerugian bagi pasien. Organisasi layanan kesehatan tersebut mencoba menyalahkan sistem diagnostik dan produsennya, tetapi hasil arbitrase menghasilkan akuntabilitas yang sepenuhnya berada di pundak penyedia layanan kesehatan, karena merekalah yang memilih dan mengkonfigurasi perangkat tersebut, dan mereka gagal melakukan tindak lanjut tambahan untuk mengkonfirmasi diagnosis.

Penerapan Akuntabilitas

Kebijakan organisasi, budaya, dan kerangka kerja tata kelola AI harus menetapkan dan menegakkan akuntabilitas yang jelas, termasuk menetapkan tanggung jawab untuk setiap aspek desain, penerapan, dan pemantauan sistem AI. Dalam beberapa kasus, organisasi harus menetapkan mekanisme untuk upaya hukum (termasuk penyelesaian sengketa) ketika terjadi dugaan kerugian. Organisasi juga perlu transparan tentang cara kerja sistem AI mereka dan menyediakan jalur bagi individu untuk menantang keputusan sistem AI.

Sangat mudah untuk menyalahkan pengembang AI dan meminta pertanggungjawaban mereka atas tindakan sistem AI mereka. Namun, jika pengembang sepenuhnya bertanggung jawab atas biaya keamanan, produk mereka mungkin menjadi kurang menarik, yang menyebabkan organisasi memilih alternatif yang lebih murah dan kurang aman. Itulah mengapa akuntabilitas pada akhirnya harus berada di pundak pihak yang menerapkan sistem tersebut. Ketika organisasi dimintai pertanggungjawaban atas perilaku sistem AI yang mereka terapkan dan gunakan, keamanan dengan cepat menjadi prioritas yang jauh lebih tinggi.

Budaya Akuntabilitas

Organisasi memiliki pandangan yang beragam mengenai topik akuntabilitas. Dengan kata lain, apakah anggota staf, manajer, pemimpin, dan eksekutif bertanggung jawab atas tindakan dan keputusan mereka? Apakah mereka diharuskan menghadapi konsekuensi ketika segala sesuatunya tidak berjalan sesuai rencana? Organisasi yang mengalami kesulitan cenderung menyalahkan pihak lain (vendor, penyedia layanan, dan sistem AI) ketika masalah terjadi. Jika sistem AI Anda membuat keputusan yang membuat organisasi Anda bermasalah, jangan salahkan sistem AI tersebut; lagipula, Anda yang memilihnya dan bertanggung jawab atas semua akibatnya.

Prinsip berpusat pada Manusia

Prinsip berpusat pada manusia menyatakan bahwa sistem AI harus dirancang dengan mempertimbangkan kesejahteraan dan martabat manusia. AI harus meningkatkan kemampuan manusia, memperbaiki kehidupan manusia, dan menjaga martabat mereka, bukan menggantikan atau mengurangi peran manusia. Sistem AI harus dirancang dan diimplementasikan selaras dengan nilai-nilai dan prioritas manusia.

Tantangan dengan Prinsip Berpusat pada Manusia

Ketika organisasi berupaya meningkatkan efisiensi dalam segala bentuk, mereka mungkin terlalu fokus pada hal-hal yang tidak penting dan memprioritaskan efisiensi finansial daripada nilai-nilai manusia. Organisasi yang menggunakan sistem AI yang berfokus pada margin keuntungan mungkin mengabaikan kesejahteraan pekerja atau pelanggan mereka. Misalnya, sebuah pabrik telah menerapkan AI untuk meningkatkan efisiensi produksi, yang mengakibatkan perpindahan pekerja. Organisasi tersebut mungkin meningkatkan posisi keuangannya tetapi dengan mengorbankan reputasinya sebagai organisasi yang berpusat pada komunitas.

Penerapan Prinsip Berpusat pada Manusia

Organisasi harus menekankan prinsip-prinsip desain yang berpusat pada manusia, memastikan bahwa sistem AI dikembangkan untuk melayani kepentingan manusia. Hal ini mungkin memerlukan penilaian dampak untuk menentukan bagaimana sistem AI meningkatkan, bukan menggantikan, kemampuan manusia. Proses pengambilan keputusan oleh sistem AI harus melibatkan manusia, khususnya di bidang jasa keuangan, perawatan kesehatan, dan peradilan.

1.5 RINGKASAN

Kecerdasan buatan mensimulasikan kecerdasan manusia dengan kemampuan seperti pembelajaran, penalaran, pemecahan masalah, persepsi, dan pemahaman bahasa. Sistem AI diklasifikasikan dalam berbagai cara, termasuk ruang lingkupnya (AI sempit/lemah, AI umum, kecerdasan buatan umum—AGI, dan kecerdasan buatan super—ASI). Model AGI dan ASI belum dikembangkan dan diumumkan kepada publik.

Fungsi AI meliputi mesin reaktif, AI memori terbatas, AI teori pikiran, AI sadar diri, dan robotika. Setiap jenis fungsi cocok untuk berbagai tugas. Beberapa teknik yang terkait dengan sistem AI meliputi pembelajaran mesin, pembelajaran mendalam, pemrosesan bahasa alami, dan sistem pakar. Teknik pelatihan sistem AI meliputi pembelajaran terawasi, tidak terawasi, semi-terawasi, dan penguatan.

Sistem AI menimbulkan risiko tertentu bagi individu, kelompok, organisasi, dan masyarakat tanpa pengamanan yang memadai. Risiko bagi individu meliputi pelanggaran privasi, bias dan diskriminasi, serangan siber, dan hilangnya otonomi dan kendali. Risiko bagi kelompok meliputi peningkatan ketidaksetaraan sosial, erosi kohesi sosial, dan serangan siber. Risiko bagi organisasi meliputi gangguan operasional, masalah hukum dan peraturan, risiko keuangan, termasuk pembengkakan biaya dan biaya hukum, serangan siber, dan kerusakan reputasi. Risiko bagi masyarakat meliputi hilangnya pekerjaan, ketidaksetaraan ekonomi, erosi kemampuan manusia untuk bertindak, serangan siber, dan manipulasi opini publik.

Organisasi sering kali merangkul dan bereksperimen dengan teknologi baru untuk mendapatkan keunggulan sebagai pelopor. Contoh historis meliputi munculnya komputer pribadi (PC), Internet, dan komputasi awan. Ada pemenang dan pecundang di setiap generasi, sebagian karena perencanaan yang baik dan sebagian karena keberuntungan.

Bagi sebagian besar organisasi, sistem AI adalah sistem yang paling kompleks dan berdampak yang pernah mereka gunakan. Kompleksitas dan kekuatan ini saja seharusnya meyakinkan setiap organisasi untuk menerapkan struktur tata kelola guna memastikan hasil yang diinginkan, karena konsekuensi dari kesalahan dapat sangat besar.

Karakteristik sistem AI, khususnya kompleksitas dan opasitas, menjadi tantangan bagi organisasi yang menggunakan sistem AI dalam pengambilan keputusan. Sistem AI dapat memberikan kecepatan dan efisiensi, tetapi mungkin memerlukan campur tangan manusia dalam peran berisiko tinggi dan berdampak besar, seperti diagnosis medis, analisis pelamar kerja, dan analisis permohonan pinjaman.

Sistem AI otonom, seperti kendaraan otonom, membutuhkan pengamanan untuk meminimalkan risiko keadaan yang berdampak besar dan tak terhindarkan. Organisasi yang sistem AI-nya memproses PII (Informasi Identitas Pribadi) harus memastikan untuk mematuhi semua hukum privasi yang berlaku, serta pemberitahuan praktik privasi organisasi sendiri. Lebih lanjut, organisasi harus melindungi data yang digunakan oleh sistem AI untuk melawan serangan siber.

AI yang bertanggung jawab adalah prinsip dasar AI. Kekuatan besar menuntut tanggung jawab besar, tetapi sistem AI juga membawa risiko kesalahan yang sangat berdampak dan godaan penyalahgunaan. Hal ini dan beberapa karakteristik AI lainnya menuntut pembentukan struktur tata kelola yang efektif untuk memastikan hanya hasil yang diinginkan.

Etika adalah prinsip inti dari AI yang bertanggung jawab, tetapi organisasi harus mendekati hal ini dengan hati-hati, karena etika dapat bersifat subjektif dan dapat diputarbalikkan sesuka hati.

Keadilan adalah bagian kunci dari AI yang bertanggung jawab, karena sangat penting bahwa sistem AI dalam peran pengambilan keputusan selalu menghasilkan hasil yang dianggap adil bagi semua pihak yang terlibat. Keamanan dan keandalan sistem AI adalah aspek lain dari AI yang bertanggung jawab, khususnya dalam konteks seperti kendaraan otonom dan operasi robotik.

Privasi dan keamanan merupakan bagian dari AI yang bertanggung jawab, di mana organisasi berkewajiban untuk mematuhi peraturan, prinsip, dan kebijakan privasi, serta melindungi semua bentuk data dalam sistem AI dari serangan siber dan bentuk penyalahgunaan lainnya.

AI yang bertanggung jawab juga mencakup transparansi dan kemampuan menjelaskan ketika sistem AI membuat keputusan yang berdampak pada pihak lain, seperti aplikasi pekerjaan dan pinjaman. Organisasi harus memahami (dan menjelaskan kepada pihak yang terkena dampak) bagaimana sistem AI menghasilkan outputnya. Beberapa jenis sistem AI lebih mudah dijelaskan daripada yang lain; organisasi harus memilih dengan bijak.

AI yang bertanggung jawab mencakup akuntabilitas organisasi tidak dapat menyalahkan sistem AI atau pembuatnya ketika segala sesuatunya tidak berjalan sesuai rencana. Organisasi perlu memiliki budaya akuntabilitas, sehingga para pengambil keputusan menganggap peran mereka dengan serius, karena mereka akan terikat oleh konsekuensi yang ditimbulkannya.

Terakhir, AI yang bertanggung jawab mencakup berpusat pada manusia. AI dapat meningkatkan efisiensi proses bisnisnya secara signifikan, tetapi dalam hal interaksi dengan karyawan dan pelanggan, tidak bijaksana untuk membuat mereka "merasa seperti angka belaka." Selain itu, sistem AI dalam peran pengambilan keputusan harus melibatkan manusia untuk memastikan keadilan bagi pihak-pihak yang terkena dampak.

BAB 2
KESIAPAN ORGANISASI

2.1 PERAN DAN TANGGUNG JAWAB PEMANGKU KEPENTINGAN TATA KELOLA AI

Sistem AI memperkenalkan kompleksitas, risiko, dan dampak sosial yang belum pernah terjadi sebelumnya ke dunia. Saat organisasi beralih dari eksperimen ke pengoperasian AI dalam skala besar, tata kelola yang kuat menjadi prasyarat, bukan hanya keunggulan kompetitif atau alat kepatuhan, tetapi keharusan moral dan strategis. Inti dari tata kelola tersebut adalah definisi peran dan tanggung jawab yang jelas. Tanpa kejelasan tentang "siapa yang bertanggung jawab atas apa," organisasi menghadapi akuntabilitas yang kabur, tekanan regulasi, dan pengawasan yang tidak efektif.

Fungsi tata kelola AI yang efektif dengan cepat dianggap sebagai bagian dari keseluruhan kehati-hatian perusahaan. Sebuah organisasi yang menghadapi masalah hukum terkait sistem AI-nya akan menghadapi konsekuensi yang lebih besar jika tidak memiliki program tata kelola AI.

Bagian ini menawarkan panduan praktis dan terperinci untuk menetapkan peran dan tanggung jawab kepada pemangku kepentingan utama dalam tata kelola AI. Bagian ini menguraikan fungsi organisasi, menawarkan pemetaan tanggung jawab berbasis siklus hidup, dan mencakup contoh dunia nyata dan strategi implementasi. Tujuannya adalah untuk memberdayakan organisasi untuk mengembangkan model tata kelola yang kuat dan mudah beradaptasi.

Tanpa peran yang didefinisikan dengan jelas, tata kelola AI gagal sebelum dimulai. Namun, ketika tanggung jawab diformalkan, dipahami, dan ditegakkan:

- Risiko dikelola dengan lebih baik.
- Komitmen etis ditegakkan.
- Inovasi berjalan dengan percaya diri dan integritas.

Bagian ini menunjukkan bahwa tata kelola AI adalah upaya tim, yang membutuhkan komitmen dari para pemangku kepentingan di berbagai tingkatan teknis, hukum, operasional, dan eksekutif. Dengan menyelaraskan orang yang tepat dengan tanggung jawab yang tepat di seluruh siklus hidup AI, organisasi dapat membangun sistem yang tidak hanya cerdas tetapi juga aman, adil, etis, dan akuntabel.

Pentingnya Peran Tata Kelola yang Terdefinisi Secara Strategis

Peran dan tanggung jawab berfungsi sebagai tulang punggung tata kelola AI. Ketika harapan didokumentasikan dan dipahami:

- Akuntabilitas dapat ditetapkan dan ditegakkan.
- Kolaborasi antar tim menjadi lebih efisien.
- Kesenjangan dan tumpang tindih dalam kepemilikan risiko berkurang.
- Sistem AI lebih selaras dengan standar bisnis, hukum, dan etika.

Tata kelola AI tidak boleh terpusat pada satu tim saja. Seperti halnya keamanan siber atau privasi data, ini adalah tanggung jawab bersama yang mencakup berbagai disiplin ilmu:

- Pemilik bisnis, data, dan produk harus menyelaraskan AI dengan tujuan perusahaan, kebutuhan pengguna, dan dampak pasar.
- Tim teknis harus memastikan ketahanan, kekokohan, dan transparansi.
- Pakar hukum dan kepatuhan harus menavigasi peraturan dan menegakkan kontrol risiko.
- Para eksekutif harus mendukung dan mendanai prioritas tata kelola.

Singkatnya, tata kelola adalah fungsi terdistribusi yang hanya berhasil jika peran-perannya didefinisikan dengan jelas.

Kategori Pemangku Kepentingan dan Peran Tata Kelolanya

Peran tata kelola AI ada di seluruh organisasi. Meskipun judulnya mungkin bervariasi, tanggung jawab cenderung berkelompok di sekitar fungsi dan tugas umum. Bagian ini memberikan gambaran umum tentang kategori utama, disertai dengan detail yang lebih lengkap.

Kepemimpinan Eksekutif

Peran umum meliputi:

- *Chief Executive Officer* (CEO)
- *Chief Operating Officer* (COO)
- *Chief Risk Officer* (CRO)
- *Chief Data or AI Officer* (CDAO)
- *Chief Compliance Officer* (CCO)
- *Chief Information Security Officer* (CISO)
- *Chief Privacy Officer* (CPO)

Tanggung jawab tata kelola dari peran-peran ini meliputi:

- Menetapkan standar dari atas untuk pengembangan dan penerapan AI yang bertanggung jawab.
- Memastikan penggunaan strategis AI organisasi selaras dengan nilai-nilai dan misi.
- Menyetujui ambang risiko dan aplikasi AI di bidang-bidang yang berisiko tinggi (misalnya, perawatan kesehatan, peradilan pidana).
- Mengalokasikan dana dan sumber daya untuk alat tata kelola, pelatihan, dan staf.
- Mempertanggungjawabkan tim atas kepatuhan terhadap protokol tata kelola.

Sebagai contoh, seorang CEO yang secara publik mendukung "Piagam Etika AI" internal dan menjadikan kepatuhan terhadapnya sebagai bagian dari KPI kepemimpinan tahunan memperkuat bahwa tata kelola bukanlah pilihan tetapi sudah tertanam.

Dewan atau Majelis Tata Kelola AI

Dewan, majelis, atau komite pengarah tata kelola AI yang umum akan mencakup para pemimpin lintas fungsi dari bidang ilmu data, keamanan siber, privasi, kepatuhan, hukum, etika, risiko, dan unit bisnis.

Tanggung jawab tata kelola meliputi:

- Mengembangkan dan memelihara kerangka kerja tata kelola AI organisasi.

- ❖ Melakukan penilaian risiko dan tinjauan untuk kasus penggunaan AI yang sensitif.
- ❖ Mengawasi implementasi audit dan penilaian dampak (misalnya, AIA, DPIA).
- ❖ Bertindak sebagai peninjau akhir untuk eskalasi risiko algoritmik.
- ❖ Memantau dan melaporkan kemajuan tujuan AI yang bertanggung jawab kepada dewan direksi.

Sebagai contoh, sebuah perusahaan ritel multinasional membentuk "Komite Pengarah AI" yang menilai apakah proyek AI yang direncanakan memenuhi kriteria etika dan peraturan yang telah ditetapkan sebelum diterapkan.

Tim Hukum, Etika, dan Kepatuhan

Tanggung jawab tata kelola penasihat hukum meliputi:

- Mengidentifikasi dan menafsirkan peraturan yang ada (misalnya, GDPR, Undang-Undang AI Uni Eropa, CCPA) dan menilai penerapannya pada organisasi dan sistem AI-nya.
- Memberikan nasihat tentang tanggung jawab hukum, kekayaan intelektual, dan risiko pihak ketiga.
- Menyusun kontrak dan SLA yang mencakup persyaratan khusus AI.
- Berpartisipasi dalam penilaian dampak dan investigasi respons insiden.

Tanggung jawab tata kelola petugas etika atau pemimpin inovasi yang bertanggung jawab meliputi:

- Menerjemahkan tujuan dan nilai-nilai organisasi ke dalam prinsip-prinsip yang dapat ditindaklanjuti untuk memilih dan merancang sistem AI.
- Memimpin penilaian dampak etika, terutama untuk sistem yang memengaruhi hak atau kebebasan.
- Menyelenggarakan "laboratorium etika" atau diskusi meja bundar untuk kasus penggunaan yang kontroversial.
- Mengembangkan narasi moral yang dapat menginformasikan pengambilan keputusan dalam kasus-kasus yang kompleks dan ambigu.

Tanggung jawab tata kelola petugas kepatuhan meliputi:

- ❖ Mengintegrasikan kontrol tata kelola ke dalam alur kerja operasional dan jalur pengiriman produk.
- ❖ Memastikan kepatuhan terhadap kebijakan internal dan standar eksternal (misalnya, ISO/IEC 42001).
- ❖ Melakukan tinjauan berkala dan audit internal pada sistem dan proses AI.
- ❖ Melatih karyawan tentang risiko kepatuhan AI dan praktik terbaik untuk mitigasinya.

Peran Kepala Petugas Etika AI (CAIEO)

Semakin banyak organisasi yang membentuk peran CAIEO untuk mengawasi tata kelola etika sistem AI mereka. Mandat mereka biasanya mencakup mempromosikan akuntabilitas publik, menerapkan strategi transparansi, dan menyeimbangkan tujuan keuntungan dan kepentingan publik.

Tim Teknis dan Rekayasa

Peran teknis dan rekayasa meliputi ilmuwan dan insinyur data, pengembang dan integrator perangkat lunak, serta rekayasa dan operasi TI. Selain para pemimpin, personel dalam tim ini tidak akan membentuk tata kelola; sebaliknya, mereka diharuskan untuk mematuhi kebijakan dan prosedur yang didorong oleh tata kelola.

Tanggung jawab ilmuwan data dan insinyur AI meliputi:

- Memilih, melatih, memvalidasi, dan mendokumentasikan model AI.
- Secara proaktif mengidentifikasi dan mengatasi masalah bias model, penyimpangan, atau perilaku yang tidak diinginkan.
- Menerapkan teknik penjelasan (misalnya, SHAP, LIME) untuk mendukung transparansi dan auditabilitas.
- Mendokumentasikan semua aspek sistem AI dan desain data.

Tanggung jawab insinyur data meliputi:

- Memastikan kualitas data, silsilah, dan keamanan di seluruh siklus hidup sistem AI.
- Implementasikan pipeline yang mendukung minimalisasi data dan kontrol privasi.
- Beri tag dan katalogkan dataset yang digunakan dalam pelatihan dan evaluasi model.

Tanggung jawab tim operasi AI meliputi:

- Mengoperasionalkan penerapan AI melalui proses SDLC/siklus hidup AI yang telah ditetapkan.
- Implementasikan pemantauan untuk pergeseran data, kinerja model, dan anomali runtime.
- Otomatiskan prosedur rollback atau peringatan sebagai respons terhadap ambang batas kritis.

Sebagai praktik terbaik, pertimbangkan untuk memperkenalkan "pos pemeriksaan tata kelola model" ke dalam siklus hidup pengembangan. Gerbang ini memeriksa dokumentasi, penilaian keadilan, dan tolok ukur kinerja sebelum mengizinkan penerapan.

Tim Keamanan Siber dan Manajemen Risiko

Tanggung jawab tim keamanan siber dan manajemen risiko meliputi:

- Mengintegrasikan risiko terkait AI ke dalam portofolio Manajemen Risiko Perusahaan (ERM).
- Mempertahankan register risiko khusus untuk sistem AI atau menerapkan tag dalam register risiko untuk mengisolasi dan menganalisis risiko khusus AI.
- Bekerja sama dengan pemimpin organisasi lain untuk mengidentifikasi dan mengevaluasi paparan keuangan, reputasi, dan hukum yang terkait dengan AI.
- Melakukan perencanaan skenario dan latihan simulasi untuk peristiwa, insiden, dan kegagalan terkait AI.
- Menentukan strategi mitigasi berdasarkan matriks kemungkinan dan tingkat keparahan. Memasukkan semua sistem AI dalam manajemen kerentanan, intelijen ancaman,
- pengerasan perangkat, pemantauan peristiwa, pengujian penetrasi, dan semua proses standar lainnya.

Misalnya, tim risiko bank mengkategorikan perangkat lunak pengenalan wajah sebagai "berisiko tinggi" karena potensi bias dan pelanggaran perlindungan data. Akibatnya, mereka mewajibkan peninjauan oleh manusia dan pemantauan dampak yang berkelanjutan.

Pemimpin Produk dan Unit Bisnis

Tanggung jawab untuk manajer produk dan pemimpin unit bisnis meliputi:

- ✓ Mendukung kasus penggunaan AI yang memenuhi tujuan bisnis dan kebutuhan pelanggan.
- ✓ Menentukan metrik keberhasilan yang selaras dengan nilai-nilai organisasi dan selera risiko.
- ✓ Memberikan masukan spesifik domain tentang relevansi data, dampak operasional, dan nuansa peraturan.
- ✓ Berpartisipasi dalam validasi model untuk memastikan output selaras dengan harapan dunia nyata.
- ✓ Bertindak sebagai penghubung antara tim teknis dan pengguna akhir.

Misalnya, di perusahaan logistik, sistem optimasi rute dimiliki bersama oleh seorang manajer produk, yang memastikan AI selaras dengan KPI pengiriman, dan seorang petugas kepatuhan, yang memastikan keadilan dalam penugasan pengemudi.

Pengguna Akhir dan Staf Garis Depan

Tanggung jawab untuk pengguna akhir dan staf meliputi:

- Berinteraksi dengan sistem AI dalam konteks operasional (misalnya, moderator konten, dokter, petugas pinjaman) dan sesuai dengan kebijakan perusahaan.
- Memantau output yang tidak akurat, tidak aman, atau bias.
- Memberikan umpan balik untuk penyempurnaan dan pelatihan ulang model.
- Sampaikan anomali atau kekhawatiran melalui saluran pelaporan tata kelola.

Aktifkan Pemberdayaan Pengguna

Pertimbangkan untuk merancang antarmuka yang memungkinkan pengguna untuk menandai rekomendasi AI, meminta intervensi manusia, atau mendapatkan penjelasan tentang hasilnya. Antarmuka ini mendorong transparansi, memberdayakan pengguna akhir, dan menanamkan tata kelola ke dalam alur kerja sehari-hari.

Pihak Eksternal

Beberapa kategori pihak eksternal berperan sebagai pemangku kepentingan dalam sistem AI suatu organisasi.

Tanggung jawab vendor dan pemasok meliputi:

- Harus memenuhi persyaratan tata kelola yang ditetapkan dalam proses pengadaan dan kontrak.
- Menyediakan dokumentasi model, penilaian risiko, dan metrik kinerja.
- Mendukung kewajiban pemantauan dan transparansi yang berkelanjutan.

Tanggung jawab regulator dan auditor meliputi:

- ➢ Meninjau kepatuhan organisasi terhadap peraturan dan standar AI yang relevan.

- Mengevaluasi sistem AI untuk keadilan, akuntabilitas, dan dampak sosial.

Masyarakat sipil, komunitas yang terdampak, dan masyarakat luas dapat berkontribusi dengan:

- Menawarkan perspektif berharga tentang bagaimana AI memengaruhi hak, kesetaraan, dan inklusi.
- Berpartisipasi dalam lokakarya desain partisipatif, kelompok fokus, atau tinjauan dampak komunitas.

Menetapkan Tanggung Jawab di Seluruh Siklus Hidup AI

Peran yang berbeda memiliki tanggung jawab yang berbeda-beda tergantung pada fase siklus hidup sistem AI. Tabel 2.1 menyediakan peta tanggung jawab siklus hidup tingkat tinggi. Tata kelola AI harus tertanam "dari kiri ke kanan" di seluruh siklus hidup. Jangan memperlakukannya sebagai hal yang dipikirkan kemudian atau sebagai tambahan. Ingatlah bahwa peran tata kelola adalah untuk memastikan hasil tertentu, yang berarti bahwa fungsi tata kelola harus ada di setiap tahap siklus hidup sistem AI.

Alat dan Teknik Implementasi

Setiap peran tata kelola harus didokumentasikan dan dikelola sesuai dengan praktik manajemen dokumen yang baik. Templat standar untuk program tata kelola AI suatu organisasi harus mendefinisikan:

- Pernyataan misi
- Hak pengambilan keputusan dan hak veto
- Prosedur eskalasi
- Metrik kinerja (misalnya, tingkat penyelesaian audit model)

Tabel 2.1

Tanggung Jawab AI di Seluruh AI Siklus Hidup

Fase	Peran Utama	Tanggung Jawab Utama
Ideasi & Desain	Eksekutif, Produk Pemilik, Etika Memimpin	Menentukan kasus penggunaan yang dapat diterima, menilai risiko etika, mengalokasikan dana.
Akuisisi Data	Insinyur Data, Legal, Kepatuhan	Pastikan pengumpulan data dilakukan secara sah, nilai bias data, dan terapkan kontrol privasi.
Model Pengembangan	Ilmuwan Data, Tim Risiko, Etika Memimpin	Melatih dan memvalidasi model, mendokumentasikan asumsi, mengevaluasi bias.
Pengujian & Validasi	Insinyur QA, Pakar Domain, Kepatuhan	Melakukan pengujian ketahanan dan keamanan, memvalidasi akurasi dan keandalan, serta menyetujui untuk penerapan.
Penyebaran	Operasi AI, TI, Pimpinan Bisnis	Memantau perilaku model, menerapkan peringatan, menegakkan SLA.
Pemantauan	Petugas Risiko, Pengguna, Kepatuhan	Lakukan audit untuk mendeteksi penyimpangan, kumpulkan umpan balik pengguna, dan perbarui dokumentasi.

Penonaktifan	Hukum, TI, Tata Kelola Papan	Hentikan penggunaan model yang sudah usang, simpan catatan data, perbarui arsip.

Tetapkan tugas tata kelola di antara para pemangku kepentingan menggunakan RACI—Bertanggung Jawab, Akuntabel, Dikonsultasikan, dan Diinformasikan. Tabel 2.2 menggambarkan matriks RACI yang umum.

Tabel 2.2

Contoh Matriks RACI atau Peran Tata Kelola dan Operasional

Tugas	Pemimpin Pengembangan	Petugas Risiko	Legal	Pemilik Produk
Pelatihan model	R	C	I	C
Evaluasi keadilan	C	R	C	C
Pengungkapan peraturan	I	C	R	C
Pemantauan pasca-penyebaran	R	A	I	C

2.2 KOLABORASI LINTAS FUNGSI DALAM PROGRAM TATA KELOLA AI

Tata kelola AI tidak beroperasi secara terisolasi; sebaliknya, ini adalah upaya lintas fungsi yang membutuhkan koordinasi di seluruh departemen, disiplin ilmu, dan bahkan pemangku kepentingan eksternal. Kompleksitas sistem AI, implikasi sosialnya, dan ketergantungan operasionalnya berarti bahwa tidak ada satu tim atau kelompok ahli pun yang dapat (atau seharusnya) mengatur AI secara efektif secara terpisah. Kolaborasi di seluruh organisasi merupakan kebutuhan strategis.

Bagian ini mengeksplorasi cara membangun dan mempertahankan kolaborasi lintas fungsi sebagai pilar program tata kelola AI Anda. Ini mencakup mengapa hal itu penting, siapa yang harus terlibat, bagaimana menyusun kolaborasi secara efektif, dan bagaimana mempertahankannya dari waktu ke waktu. Anda juga akan menemukan contoh dan alat untuk membantu mengoperasionalkan prinsip-prinsip ini dalam organisasi Anda.

Kolaborasi lintas fungsi bukanlah sekadar formalitas: ini adalah penghubung yang memastikan sistem AI tetap selaras dengan budaya perusahaan, nilai-nilai kemanusiaan, tujuan bisnis, peraturan, dan harapan masyarakat.

Mengapa Kolaborasi Lintas Fungsi Sangat Penting

Karena kompleksitasnya dan cara-cara baru dalam mengonsumsi data, menghasilkan data, dan memengaruhi hasil bisnis, sistem AI memengaruhi:

- Risiko hukum dan peraturan (misalnya, diskriminasi, pelanggaran privasi data)
- Risiko reputasi (misalnya, penggunaan yang tidak etis, pengambilan keputusan yang tidak transparan)
- Ketahanan teknis (misalnya, bias model, penyimpangan, serangan musuh)
- Operasi bisnis (misalnya, efisiensi otomatisasi, skalabilitas)
- Dampak terhadap manusia (misalnya, etika, keadilan, martabat, kehilangan pekerjaan)

Tidak ada satu tim pun—baik tim hukum, ilmu data, TI, keamanan siber, manajemen risiko, atau SDM—yang dapat sepenuhnya menilai atau mengurangi semua risiko ini. Kolaborasi lintas fungsi memastikan bahwa berbagai perspektif diterapkan pada pengawasan AI.

Ketika tata kelola AI terpusat di bawah satu fungsi perusahaan, seringkali kurang relevansi bisnis atau kedalaman teknis. Kepemilikan bersama di seluruh organisasi memastikan:

- Kepatuhan yang lebih tinggi terhadap proses tata kelola
- Pengambilan keputusan yang lebih peka terhadap konteks
- Dukungan organisasi yang lebih besar
- Budaya yang memprioritaskan inovasi yang bertanggung jawab
- Risiko masalah dan insiden pasca-implementasi yang lebih rendah

Mengapa Kita Tidak Bisa Akur?

Banyak organisasi mengalami gesekan antar departemen, seperti TI dan keamanan siber, yang seringkali tampak bertentangan. Organisasi yang ingin sukses dengan AI harus menemukan cara untuk mengesampingkan perbedaan mereka dan bergabung untuk mencapai hasil terbaik yang didukung AI.

Prinsip-prinsip Tata Kelola Lintas Fungsi yang Efektif

Untuk membangun kolaborasi yang tahan lama dan efektif, organisasi harus menanamkan pendorong struktural, budaya, dan prosedural ke dalam hubungan dan proses mereka. Keragaman keahlian adalah salah satu pendorong struktural tersebut, dengan melibatkan pemangku kepentingan dengan latar belakang yang beragam, termasuk:

- Peran teknis: Ilmuwan data, insinyur AI, arsitek TI
- Peran hukum/kepatuhan: Penasihat privasi, keamanan siber, pakar regulasi
- Peran etika: Kepala petugas kepatuhan, pemimpin AI yang bertanggung jawab
- Pemimpin bisnis: Pemilik produk, manajer proses
- Staf operasional: SDM, keuangan, layanan pelanggan
- Perwakilan pengguna akhir: Pengguna internal dan eksternal

Beragam Perspektif Meningkatkan Deteksi Risiko

Dalam satu kasus, sebuah model bahasa ditandai sebagai bias oleh seorang ahli bahasa, bukan oleh seorang ilmuwan data, yang mengidentifikasi stereotip sosiolinguistik selama pengujian. Tim peninjau multidisiplin memperluas jangkauan deteksi risiko.

Meskipun kolaborasi membutuhkan fleksibilitas, definisi peran yang jelas membantu mencegah kebuntuan dan perebutan kekuasaan. Definisikan:

- Siapa yang bertanggung jawab atas keputusan mana (misalnya, persetujuan implementasi akhir)
- Siapa yang memberikan masukan (misalnya, etika, hukum, keamanan siber)
- Siapa yang bertanggung jawab atas dokumentasi?
- Siapa yang memantau hasil pasca-implementasi?

Pertimbangkan untuk menggunakan matriks RACI atau kerangka kerja hak pengambilan keputusan untuk mengurangi ambiguitas. Simpan catatan bisnis yang komprehensif, termasuk risalah rapat dan keputusan yang dapat ditinjau kemudian.

Organisasi tidak boleh memperlakukan kolaborasi sebagai sesuatu yang episodik (misalnya, hanya selama tinjauan risiko). Sebaliknya, tanamkan kolaborasi ke dalam siklus hidup AI sebagai fungsi iteratif:

- Tinjauan desain awal
- Perencanaan akuisisi data
- Tahap validasi model
- Pemeriksaan kesiapan implementasi
- Simulasi respons insiden

Titik pemeriksaan lintas fungsi dan gerbang keputusan di seluruh proses pengembangan sistem AI menciptakan hasil yang lebih aman dan selaras.

Membangun Struktur Kolaboratif

Jika Anda bertanya kepada sepuluh orang apa yang dimaksud dengan kolaborasi, Anda mungkin akan mendapatkan beberapa jawaban yang berbeda. Berikut adalah beberapa ide yang dapat dipertimbangkan untuk membuat orang-orang dari seluruh organisasi bekerja sama dalam masalah terkait AI.

Kelompok Kerja dan Dewan Penasihat

Kelompok kerja tata kelola AI lintas fungsi yang tetap dapat berfungsi sebagai pusat koordinasi dan kolaborasi yang berkelanjutan. Pertimbangkan keanggotaan untuk:

- ✓ Perwakilan dari bidang hukum, risiko, dan kepatuhan
- ✓ Pemimpin teknis dari AI/ML atau rekayasa
- ✓ Perwakilan produk/bisnis, CCO atau penasihat etika
- ✓ Perwakilan keamanan siber dan privasi
- ✓ Penghubung pengguna akhir atau komunitas (jika berlaku)

Tanggung Jawab:

- Meninjau kasus penggunaan berisiko tinggi
- Mengkoordinasikan penilaian risiko (AIA, DPIA, kartu model)
- Memperbaiki kebijakan tata kelola
- Meningkatkan kekhawatiran ke dewan AI tingkat eksekutif
- Melacak metrik akuntabilitas dan pelatihan

Misalnya, lembaga keuangan membentuk "Dewan AI Bertanggung Jawab" yang terdiri dari 12 anggota yang berganti-ganti dari berbagai departemen. Dewan tersebut bertemu setiap bulan untuk menilai model yang sedang dikembangkan dan setiap tahun untuk menyarankan pembaruan pada piagam tata kelola. Secara opsional, dewan tersebut dapat diberi wewenang untuk membuat perubahan pada piagam tata kelola, bukan hanya sekadar memberikan saran.

Menetapkan Tinjauan Kasus Penggunaan Lintas Fungsi

Menetapkan protokol tinjauan untuk setiap sistem AI yang melampaui ambang risiko tertentu atau diklasifikasikan dalam kategori risiko yang lebih tinggi. Tujuan dari tinjauan ini meliputi:

- Menilai risiko etika dan hukum
- Memvalidasi representasi pemangku kepentingan
- Menguji kinerja model di berbagai kelompok demografis
- Mengkonfirmasi rencana mitigasi

Peserta yang dibutuhkan meliputi:

- ❖ Pemimpin ilmu data
- ❖ Petugas risiko/kepatuhan
- ❖ Pemilik bisnis
- ❖ Peninjau etika
- ❖ Advokat pengguna akhir

Kapan tinjauan harus dilakukan? Pertimbangkan tinjauan kasus penggunaan ketika:

- Keputusan otomatis memengaruhi hak individu
- Data mencakup data pribadi atau biometrik yang sensitif
- Sistem berdampak pada perekrutan, pemberian pinjaman, atau perawatan medis
- Kompleksitas model mengurangi kemampuan untuk dijelaskan
- Peraturan yang diperluas meningkatkan pengawasan regulasi

Templat dan Artefak Kolaborasi

Organisasi dapat menstandarisasi dokumentasi dan catatan pengambilan keputusan mereka dengan artefak bersama:

- **Lembar fakta model atau kartu model:** Mencakup tujuan, keterbatasan, dan metrik evaluasi
- **Formulir penilaian risiko:** Panduan terstruktur untuk kategori risiko (bias, penyimpangan, penyalahgunaan)
- **Penilaian dampak:** Templat untuk Penilaian Dampak Perlindungan Data (DPIA) dan Penilaian Dampak Algoritma (AIA)
- **Daftar periksa etika:** Mencakup martabat, keadilan, penjelasan, interpretasi, dan pengawasan manusia

Dokumen bersama mengurangi kesalahpahaman dan berfungsi sebagai catatan kolaborasi. Sama seperti sistem AI yang harus memiliki tingkat transparansi yang tinggi, demikian pula tata kelola AI.

Membina Budaya Kolaboratif

Bahkan dengan struktur tata kelola yang ada, kolaborasi dapat gagal tanpa akuntabilitas dan penguatan budaya. Tim teknis dan non-teknis seringkali berbicara dengan "bahasa" yang berbeda. Organisasi dapat mengatasi hal ini dengan:

- ✓ Lokakarya bersama untuk membangun kosakata bersama (misalnya, apa arti "etis" dan "keadilan" dalam konteksnya?)
- ✓ Program magang antar departemen
- ✓ Peran tata kelola rotasi (misalnya, penasihat hukum yang ditempatkan bersama tim data selama satu kuartal)

Sebagai contoh, sebuah perusahaan teknologi menyelenggarakan “AI Ethics Hackathon” tahunan di mana para insinyur, desainer, pengacara, dan ilmuwan sosial berkolaborasi untuk

mengembangkan solusi terhadap tantangan dan jebakan AI yang sudah dikenal. Tata kelola sering diperlakukan sebagai pekerjaan "tambahan". Untuk menghargai kontribusi lintas tim, ubah sikap ini dengan:

- Memasukkan partisipasi tata kelola dalam tinjauan kinerja
- Membuat program pengakuan untuk kolaborasi
- Memberikan penghargaan atas upaya tata kelola lintas fungsi yang berhasil

Tata Kelola sebagai KPI

Sebuah perusahaan melacak "Mitigasi risiko proyek AI sebelum penerapan" sebagai indikator kinerja utama (KPI) untuk tim teknis dan hukum. Pelacakan ini menciptakan insentif bersama.

Kolaborasi seringkali gagal ketika suara-suara tertentu mendominasi percakapan. Ciptakan keamanan psikologis di tengah pusaran dinamika kekuasaan dengan:

- Mengganti fasilitator rapat
- Menciptakan saluran masukan anonim
- Memberdayakan staf junior untuk menyampaikan kekhawatiran etis
- Membutuhkan partisipasi panel yang beragam dalam pengambilan keputusan

Kiat Wawasan Etis

Orang-orang yang paling dekat dengan teknologi sering kali melihat risiko sejak dini, jadi pastikan untuk memberi mereka kesempatan untuk berbicara. Dorong masukan dari mereka yang memiliki kekuasaan institusional lebih rendah tetapi wawasan yang lebih berharga.

Kolaborasi Eksternal dan Lintas Organisasi

Dalam konteks teknologi informasi (yang mencakup AI), organisasi tidak lagi terisolasi; sebaliknya, mereka bekerja sama erat dengan vendor dan penyedia layanan pihak ketiga. Saat menggunakan sistem AI pihak ketiga:

- Membutuhkan partisipasi dalam tinjauan tata kelola Anda
- Meninjau dokumentasi model dan penilaian risiko mereka
- Menyelaraskan prinsip bersama (misalnya, tidak menggunakan atribut sensitif untuk prediksi)
- Menetapkan protokol respons insiden bersama

Misalnya, sebuah jaringan ritel menggunakan vendor pengenalan wajah pihak ketiga dan membutuhkan pengujian bersama untuk bias demografis sebelum penerapan.

Dalam konteks regulasi, organisasi perlu terlibat secara proaktif dengan:

- Regulator untuk menyelaraskan praktik mereka dengan hukum yang berkembang (misalnya, Undang-Undang AI Uni Eropa)

- Konsorsium industri (misalnya, IEEE, ISO, NIST, ENISA) untuk berkontribusi pada penetapan standar
- Lembaga akademis untuk perspektif kritis dan metodologi pengujian

Misalnya, sebuah perusahaan asuransi bermitra dengan pusat etika universitas lokal untuk bersama-sama merancang tolok ukur keadilan untuk model penjaminan.

Dalam aplikasi berdampak tinggi (misalnya, penegakan hukum, layanan sosial), kolaborasi komunitas sangat penting. Organisasi dapat berkolaborasi dengan:

- Melakukan kelompok fokus dan wawancara pemangku kepentingan
- Mengundang kelompok masyarakat sipil untuk meninjau kasus penggunaan
- Membuat lokakarya desain partisipatif
- Menerbitkan pernyataan dampak dan meminta umpan balik

Ingat, lintas fungsi tidak hanya berarti lintas departemen. Ini berarti mendengarkan mereka yang terpengaruh oleh sistem AI, terutama ketika keputusan memiliki dampak signifikan pada bidang-bidang seperti perumahan, kepolisian, atau kredit.

Mempertahankan Kolaborasi dari Waktu ke Waktu

Antusiasme awal sering memudar ketika masa bulan madu berakhir. Kolaborasi tata kelola AI yang berkelanjutan membutuhkan:

- ❖ Piagam tata kelola: Mengkodifikasi peran, ritme pertemuan, dan jalur eskalasi
- ❖ Program pelatihan: Melatih tim lintas fungsi bersama tentang risiko AI, bias, dan pembaruan hukum
- ❖ Metrik bersama: Melacak bagaimana kolaborasi meningkatkan mitigasi risiko, transparansi, dan kepatuhan
- ❖ Retrospektif: Mengadakan postmortem atas kegagalan tata kelola atau hampir gagal untuk meningkatkan

Tata kelola tidak berakhir ketika sistem AI diterapkan. Sebaliknya, ini baru permulaan.

Sebagai contoh, tim AI perawatan kesehatan mengadakan retrospektif tata kelola triwulanan, termasuk staf klinis, untuk menilai konsekuensi yang tidak diinginkan dari model yang diterapkan.

2.3 PELATIHAN DAN KESADARAN TERMINOLOGI, STRATEGI, DAN TATA KELOLA AI

Kerangka kerja tata kelola AI yang dirancang dengan baik cenderung gagal tanpa pemahaman dan dukungan organisasi. Dari ruang rapat hingga ruang server, para pemangku kepentingan membutuhkan kosakata bersama dan pemahaman umum tentang apa itu AI, bagaimana AI selaras dengan tujuan strategis organisasi, dan bagaimana AI harus diatur untuk memastikan keberhasilan. Keselarasan ini dimulai dengan pendidikan.

Program pelatihan dan kesadaran yang efektif adalah kunci utama yang mengubah kebijakan tata kelola dari dokumen menjadi perilaku organisasi sehari-hari. Pelatihan dan kesadaran bukanlah tambahan untuk tata kelola AI melainkan mekanisme yang mengubah kebijakan menjadi praktik. Dengan membekali para pemangku kepentingan dengan pengetahuan, kosakata, dan jalur untuk bertindak secara bertanggung jawab, organisasi mengurangi risiko dan meningkatkan keselarasan. Hal ini membangun kefasihan, memperkuat

akuntabilitas, dan memungkinkan tim untuk mengidentifikasi risiko sebelum risiko tersebut berujung pada kegagalan.

Bagian ini mengeksplorasi cara merancang dan mengimplementasikan inisiatif pelatihan dan kesadaran tata kelola AI yang sukses. Bagian ini membahas segmentasi audiens, pengembangan kurikulum, strategi penyampaian, dan metrik keberhasilan.

Manfaatkan Apa yang Sudah Anda Miliki

Seperti yang dibahas di seluruh buku ini, Anda perlu memanfaatkan proses, sumber daya, dan teknik yang sudah ada saat membangun tata kelola AI. Di sini, Anda perlu menggunakan sumber daya dan metode yang digunakan untuk memberikan pelatihan tata kelola AI, seperti sistem manajemen pembelajaran (LMS), pencatatan, kebijakan, dan pesan eksekutif.

Mengapa Pelatihan dan Kesadaran Sangat Penting

AI bukanlah satu teknologi tunggal melainkan serangkaian metode, alat, dan filosofi yang memengaruhi personel dan proses dengan cara baru. Di antara para pemangku kepentingan, literasi AI akan berkisar dari pemula hingga ahli. Misalnya, tim hukum mungkin tidak memahami overfitting, pengembang mungkin tidak memahami batasan peraturan, dan eksekutif mungkin tidak fasih dalam persyaratan tata kelola.

Pelatihan menjembatani kesenjangan ini dengan:

- Mendefinisikan batasan dan struktur pengembangan sistem
- Menciptakan bahasa bersama Menyelaraskan prioritas di berbagai peran
- Mengurangi risiko melalui deteksi dini
- Memberdayakan tim untuk bertindak dengan percaya diri dalam batasan tata kelola

Tanpa Kesadaran dan Akuntabilitas, Kebijakan = Sandiwara

Suatu organisasi mungkin memiliki pedoman etika atau protokol risiko, tetapi jika karyawan tidak memahaminya atau tidak tahu bahwa pedoman atau protokol tersebut ada mereka cenderung tidak akan mengikutinya.

Cakupan Program Pelatihan

Tidak setiap pemangku kepentingan perlu menjadi ilmuwan data, dan tidak setiap ilmuwan data perlu menjadi manajer risiko; namun, setiap peran harus cukup memahami AI untuk membuat keputusan yang tepat, etis, dan sesuai. Tujuan inti harus mencakup:

- **Terminologi AI:** Konsep-konsep kunci (misalnya, pembelajaran mesin, jaringan saraf, bias algoritmik, pergeseran model, kemampuan menjelaskan)
- **Strategi AI:** Bagaimana AI mendukung misi, prioritas, dan postur risiko organisasi, dalam bentuk prinsip atau contoh singkat
- **Prinsip tata kelola:** Kebijakan, peran, proses peninjauan, persyaratan dokumentasi
- **Kesadaran risiko:** Risiko etis, hukum, operasional, dan reputasi yang terkait dengan AI

- **Saluran eskalasi:** Cara menyampaikan kekhawatiran, menandai anomali, atau mencari panduan

Pelatihan harus disesuaikan dengan kebutuhan kelompok pemangku kepentingan yang berbeda. Pertimbangkan untuk membuat "persona" audiens dan menyelaraskan tujuan pembelajaran sesuai dengan itu. Tabel 2.3 memberikan beberapa contoh yang dapat Anda pertimbangkan untuk digunakan sebagai titik awal.

Tabel 2.3
Memetakan Audiens ke Kebutuhan dan Topik Pelatihan

Audience	Kebutuhan	Contoh Topik
Eksekutif & Anggota Dewan	Penyelarasan strategis, paparan hukum, risiko tingkat perusahaan	Strategi AI, tren regulasi, risiko reputasi
Pengembang & Ilmuwan Data	Ketelitian teknis, alat keadilan, standar	Dokumentasi model, mitigasi bias, reproduksibilitas, vendor dan alat standar
Hukum, Risiko, dan Kepatuhan	Pemahaman regulasi, kesiapan audit.	Hukum terkait AI, penilaian dampak, risiko vendor, harapan dan persyaratan pelanggan.
Bisnis/Produk Tim	Penyelarasan nilai, kesadaran dampak	Verifikasi kasus penggunaan, kriteria keberhasilan, dampak bagi pengguna.
Pengguna Garis Depan	Pemahaman operasional	Perilaku sistem, alur kerja dengan campur tangan manusia, opsi penyelesaian masalah.
TI dan Keamanan Siber	Kontrol, operasional harian, manajemen insiden	Pemantauan sistem dan peristiwa, respons terhadap peristiwa dan insiden, tanggung jawab pemilik kendali.
Semua Staf	Literasi dasar, budaya etis	Definisi, contoh, jalur eskalasi

Struktur Kurikulum

Bagian ini memberikan contoh struktur kurikulum yang dapat disesuaikan dengan kebutuhan organisasi tertentu. Jenis pertimbangan yang dapat memengaruhi struktur ini meliputi ukuran organisasi, sektor pasar, lanskap peraturan, selera risiko, dan budaya perusahaan.

Contoh Struktur dan Area Konten

Tabel 2.4 berisi contoh struktur untuk berbagai topik dalam pelatihan tata kelola AI. Berdasarkan sektor bisnis organisasi Anda, lanskap peraturan, dan faktor relevan lainnya, Anda perlu menyesuaikan pendekatan ini untuk memenuhi kebutuhan spesifik organisasi Anda.

Tabel 2.4
Memetakan Audiens ke Kebutuhan dan Topik Pelatihan

Topik Pelatihan	Cakupan Subjek
Dasar-dasar AI (untuk semua)	Apa itu AI, ML, dan otomatisasi? Bagaimana AI berbeda dari perangkat lunak tradisional?

	Mitos dan kesalahpahaman umum
Penyelarasan strategis (untuk manajer, eksekutif)	Mengapa organisasi tersebut berinvestasi di AI? Bagaimana AI akan memungkinkan dan memajukan tujuan perusahaan Peran tata kelola dalam meningkatkan skala AI secara aman. Nilai bisnis dan risiko reputasi
Kerangka tata kelola (untuk semua)	Ikhtisar kebijakan dan peran internal Dokumentasi apa yang diperlukan Bagaimana model ditinjau dan disetujui
Etika dan risiko (untuk semua)	Prinsip-prinsip: keadilan, akuntabilitas, transparansi Contoh jebakan etika dan cara mengatasinya Proses respons insiden
Aplikasi berbasis peran (jalur yang disesuaikan)	Studi kasus yang relevan dengan masing-masing departemen Panduan langkah demi langkah (misalnya, "Cara melakukan tinjauan keadilan")

Contoh Judul Modul

Berikut beberapa saran untuk penamaan modul pelatihan. Meskipun mudah untuk terlalu memikirkan nama modul dan kursus pelatihan, penting juga agar nama modul sesuai dengan materi pokoknya. Menambahkan sedikit daya tarik juga bermanfaat, karena dapat memicu rasa ingin tahu peserta pelatihan.

- ❖ “AI vs. Perangkat Lunak Tradisional: Mengapa Tata Kelola Khusus AI Diperlukan”
- ❖ “Apa yang Perlu Diketahui Setiap Manajer tentang Bias Algoritma”
- ❖ “Di Dalam Kartu Model: Transparansi dalam Praktik”
- ❖ “Eskalasi Risiko dalam Siklus Hidup AI: Kapan Harus Mengangkat Tangan Anda”
- ❖ “Tata Kelola dalam Aksi: Tinjauan Kasus Penggunaan Lintas Fungsi”

Menyampaikan Pelatihan

Metode penyampaian pelatihan tata kelola AI sangat beragam, seperti halnya audiens, dalam hal peran dan gaya belajar mereka.

Pembelajaran Campuran

Penting untuk menyadari bahwa tidak semua orang belajar dengan cara yang sama; sebaliknya, setiap individu memiliki gaya belajar yang berbeda. Bijaksana untuk menggunakan kombinasi format penyampaian untuk memenuhi kebutuhan beragam pembelajar dan memperkuat materi. Tabel 2.5 mencakup gaya belajar yang umum.

Tabel 2.5

Gaya Belajar dan Kekuatannya

Format	**Kekuatan**	**Contoh**
Lokakarya langsung	Interaktif, membangun tim	Simulasi kasus lintas fungsi
modul e-Pembelajaran	Dapat diskalakan, asinkron	Kursus orientasi karyawan baru, penyegaran kebijakan.
Panduan kerja dan buku petunjuk	Spesifik peran, dapat diakses	Daftar periksa etika, eskalasi diagram alur

Pembelajaran sejawat	Kaya konteks, sosial	Sesi makan siang sambil belajar, komunitas praktik.
Pembelajaran mikro	Hemat waktu, menarik	Video lima menit, kuis singkat

Konten Wajib vs. Pilihan

Mengapa hanya menyediakan modul pelatihan wajib ketika beberapa personel memiliki keinginan untuk belajar lebih banyak? Anda perlu menetapkan modul wajib inti (misalnya, untuk semua karyawan setiap tahun) dan modul pilihan untuk peran khusus. Jalur pilihan dapat mencakup:

- "Kemampuan AI yang Sedang Berkembang" untuk semua
- "Kepatuhan AI dan Hukum" untuk tim hukum
- "Mengoperasionalkan Tata Kelola dalam Operasi AI" untuk insinyur
- "Konsep Desain AI yang Berpusat pada Manusia" untuk tim produk

Pertimbangan Regional dan Budaya

Organisasi multinasional dan multikultural mungkin perlu mempertimbangkan untuk menyesuaikan pelatihan tata kelola AI mereka agar mencerminkan:

- Peraturan lokal (misalnya, CCPA di California, PIPEDA di Kanada, GDPR di Eropa)
- Preferensi bahasa
- Interpretasi budaya tentang keadilan, bias, privasi, dan risiko

Memelihara Catatan Pelatihan

Melacak kehadiran dan skor kuis kompetensi adalah praktik penting untuk pelatihan perusahaan di semua bidang fokus. Organisasi dengan budaya akuntabilitas yang sehat dapat melacak kehadiran dan menindaklanjuti karyawan yang tidak hadir. Melacak skor kuis memberikan bukti tambahan tidak hanya tentang kehadiran tetapi juga pemahaman, mengurangi atau menghilangkan kemampuan peserta untuk mengklaim bahwa mereka tidak memahami pelatihan.

Beberapa peraturan, seperti Undang-Undang AI Uni Eropa, secara eksplisit mensyaratkan dokumentasi pelatihan karyawan dalam peran terkait AI. Sementara itu, pelanggan, investor, dan publik semakin mengevaluasi perusahaan berdasarkan kemampuan mereka untuk bertindak secara bertanggung jawab dengan AI.

Membangun Kesadaran di Luar Pelatihan Formal

Organisasi harus memperlakukan tata kelola AI seperti sebuah merek—ia membutuhkan visibilitas dan penguatan. Ide-ide tersebut meliputi:

- Buletin bulanan tentang "AI Bertanggung Jawab dalam Praktik"
- Blog internal yang menampilkan wawancara dengan para juara tata kelola AI
- "Tips mingguan" tata kelola di saluran Teams
- Kontes trivia tata kelola AI
- Tantangan "Temukan bias"
- Simulasi kegagalan tata kelola ala escape room

Misalnya, sebuah perusahaan logistik global menerbitkan "Ringkasan Keselamatan" triwulanan yang kadang-kadang menyertakan konten terkait AI, seperti insiden model, pembaruan peraturan, dan kisah sukses dari audit internal.

Budaya Dimulai dari Puncak

Keterlibatan eksekutif dalam topik tata kelola AI membantu memberi sinyal bahwa tata kelola AI sangat penting bagi keberhasilan organisasi, bukan sekadar formalitas kepatuhan.

Melacak dan Mengukur Keberhasilan

Organisasi yang berorientasi pada metrik tidak akan berhenti pada pelacakan tingkat penyelesaian, tetapi juga akan berupaya mengukur dampak. Beberapa KPI yang berguna meliputi:

- Persentase staf yang menyelesaikan pelatihan dalam 90 hari
- Penilaian pra/pasca literasi AI
- Persentase proyek AI dengan dokumentasi model lengkap tentang eskalasi risiko yang diprakarsai oleh staf terlatih
- Kepuasan pengguna terhadap alat dan saluran tata kelola

Dalam semangat peningkatan berkelanjutan, organisasi harus membangun lingkaran umpan balik formal. Contohnya meliputi:

- Umpan balik peserta didik tentang kejelasan dan relevansi
- Saran untuk topik modul baru
- Penerapan dan kekhawatiran departemen

Misalnya, sebuah organisasi memperhatikan kebingungan berulang seputar konsep "kemampuan menjelaskan". Sebagai tanggapan, tim tata kelola AI membuat lokakarya tentang interpretasi model, memberikan contoh bagaimana organisasi menangani topik ini.

Kesalahan Umum dan Cara Menghindarinya

Perencana yang bijak memahami bahwa belajar dari kesalahan orang lain itu penting. Organisasi yang tidak ingin memulai dari awal lagi dengan pelatihan tata kelola AI lebih memilih untuk melakukan hal yang benar sejak awal. Tabel 2.6 mencakup beberapa contoh.

Tabel 2.6

Menghindari Pita Pelatihan

Jebakan	**Mitigasi**
Konten yang berlaku untuk semua	Segmentasikan pelatihan berdasarkan peran dan konteks.
Modul yang terlalu teknis	Gunakan bahasa dan visual yang sederhana
Kurangnya partisipasi eksekutif	Wajibkan kehadiran eksekutif tingkat C (CEO, CFO, CFO) dalam sesi-sesi penting.
Tidak ada tindak lanjut	Mengintegrasikan hasil pelatihan ke dalam tujuan kinerja dan proses audit.
Pelatihan sebagai acara sekali waktu	Buat jalur pembelajaran berkelanjutan dan siklus penyegaran.

Mempertahankan Program dari Waktu ke Waktu

Pelatihan tata kelola AI bukanlah acara sekali jadi di awal kerja sama organisasi dengan AI. Sebaliknya, organisasi harus mengembangkan lingkungan pengetahuan tata kelola AI yang dinamis yang mencakup:

- ✓ Portal pusat dengan alat, templat, dan kebijakan
- ✓ Perpustakaan lokakarya sebelumnya yang dapat diakses sesuai permintaan
- ✓ Forum internal untuk diskusi dan tanya jawab

Konten pelatihan perlu diperbarui secara berkala, terutama ketika terjadi perubahan peraturan, ketika kemampuan AI yang baru muncul tersedia, dan setelah insiden terkait AI terjadi di dalam organisasi dan di tempat lain.

Organisasi yang serius tentang pengembangan staf dan kepemimpinan dapat mempertimbangkan untuk menghubungkan pelatihan tata kelola AI dengan rencana pengembangan profesional, metrik kinerja, dan jenjang karier. Misalnya, kebijakan SDM suatu organisasi mencakup "Kemahiran Tata Kelola AI" sebagai kompetensi inti untuk promosi ke kepemimpinan tim.

2.4 MENYESUAIKAN TATA KELOLA AI DENGAN KONTEKS ORGANISASI

Tata kelola AI bukanlah solusi yang cocok untuk semua. Jika demikian, kita tidak memerlukan buku ini atau sertifikasi AIGP kita hanya akan mengambil kerangka kerja dan menjadikannya milik kita sendiri. Sama seperti tidak ada dua organisasi yang membangun, menggunakan, atau bergantung pada AI dengan cara yang sama, pendekatan mereka terhadap tata kelola AI harus mencerminkan konteks unik mereka. Bank multinasional yang menerapkan model risiko prediktif akan membutuhkan program tata kelola yang pada dasarnya berbeda dari perusahaan rintisan teknologi yang mengoptimalkan logistik.

Bagian ini menguraikan cara membedakan strategi tata kelola AI berdasarkan enam faktor penting:

- Ukuran perusahaan
- Kematangan organisasi
- Industri
- Produk dan layanan
- Tujuan strategis
- Toleransi risiko

Baik Anda membantu perusahaan kecil untuk memulai atau mengevaluasi posisi perusahaan global, memahami dimensi-dimensi ini akan memandu desain tata kelola AI yang tepat, terukur, dan berkelanjutan.

Tata kelola AI harus peka terhadap konteks. Perusahaan rintisan tidak boleh mengadopsi kontrol yang sama dengan bank; perusahaan logistik tidak memerlukan proses yang sama dengan lembaga pemerintah. Agar efektif, tata kelola harus proporsional, relevan, dan adaptif.

Bagian ini mengilustrasikan bagaimana strategi tata kelola bervariasi menurut ukuran perusahaan, kematangan, industri, risiko produk, tujuan, dan selera risiko. Memahami dan

menyesuaikan diri dengan faktor-faktor ini memastikan bahwa kerangka kerja tata kelola tidak hanya sesuai tetapi juga praktis, dapat dicapai, berkelanjutan, dan selaras dengan bisnis.

Ukuran Perusahaan: Menskalakan Tata Kelola ke Jejak Organisasi

Organisasi yang lebih kecil beroperasi jauh berbeda dari organisasi yang lebih besar. Proses yang dibutuhkan oleh organisasi yang lebih besar, yang melibatkan tindakan puluhan, ratusan, atau ribuan karyawan, akan memberatkan bagi organisasi yang lebih kecil. Demikian pula, kesederhanaan dan keanggunan proses organisasi kecil akan menjadi kacau jika diadopsi oleh organisasi besar.

Bagian ini membagi tata kelola menjadi tiga segmen: organisasi kecil dengan hingga 200 karyawan, organisasi menengah dengan 200–5.000 karyawan, dan perusahaan dengan 5.000 karyawan atau lebih. Isi bagian ini berfungsi sebagai pedoman. Seperti aspek lain dari tata kelola AI, penting untuk menerapkan kebijakan, proses, dan catatan dengan cara yang konsisten dengan praktik yang ada.

Perusahaan Kecil dan Startup (1–200 Karyawan)

Ciri-ciri Umum:

- Sumber daya hukum dan kepatuhan yang terbatas
- Sedikit, atau bahkan tidak ada, struktur tata kelola formal
- Struktur organisasi yang datar
- Proses informal dan kelincahan yang tinggi
- Pemberdayaan karyawan yang tinggi
- Rasio inovasi-ke-tata kelola yang tinggi

Rekomendasi Tata Kelola:

- **Mulailah dengan prinsip-prinsip inti:** Piagam etika AI yang ringkas dapat memberikan panduan sebelum kebijakan formal ditetapkan.
- **Tunjuk seorang pemimpin tata kelola:** Bahkan tanpa komite penuh, tetapkan satu atau dua pemimpin lintas fungsi.
- **Gunakan alat yang ringan:** Spreadsheet, daftar periksa bersama, dan tinjauan risiko manual adalah titik awal yang praktis.
- **Alih daya jika diperlukan:** Tinjauan hukum, audit, atau pengujian risiko dapat dikontrakkan kepada perusahaan eksternal.

Contoh startup edtech dengan 30 karyawan memasukkan pertanyaan risiko etika ke dalam rapat desain produknya, dipandu oleh daftar periksa AI yang bertanggung jawab dalam satu halaman.

Organisasi Menengah (200–5.000 Karyawan)

Ciri-ciri Umum:

- Beberapa peran dan proses tata kelola yang diformalkan
- Kebijakan yang terdokumentasi
- Fungsi kepatuhan internal tahap awal
- Ekspansi aktif AI/ML di berbagai fungsi

Rekomendasi Tata Kelola:

- Bentuk dewan tata kelola AI: Sertakan perwakilan dari TI, hukum, produk, dan ilmu data.
- Tetapkan ambang batas tinjauan berbasis risiko: Wajibkan penilaian dampak untuk model dengan keputusan yang berhadapan langsung dengan pelanggan atau keputusan berdampak tinggi.
- Mulai integrasi siklus hidup: Perkenalkan titik pemeriksaan dokumentasi selama pengembangan dan penerapan.

Sebagai contoh, sebuah jaringan ritel dengan 2.000 karyawan membutuhkan templat "Proposal Penggunaan AI" untuk setiap inisiatif AI baru, yang ditinjau setiap bulan oleh kelompok pengarah.

Perusahaan Besar (5.000+ Karyawan)

Ciri-ciri Umum:

- Aplikasi AI yang luas di berbagai departemen
- Departemen TI terpisah untuk setiap divisi atau unit bisnis
- Struktur tata kelola formal yang sudah ada
- Profil risiko regulasi dan reputasi yang kompleks
- Pemberdayaan karyawan yang terdefinisi
- Tim hukum, risiko, dan tata kelola data yang berdedikasi

Rekomendasi Tata Kelola:

- ✓ Formalkan tata kelola multi-lapisan: Gunakan model federasi dengan kebijakan terpusat dan akuntabilitas terdistribusi.
- ✓ Integrasikan tata kelola dengan ERM, keamanan siber, dan tata kelola TI: Selaraskan pengawasan AI dengan proses keamanan siber, pengadaan, dan audit.
- ✓ Investasikan pada perangkat dan otomatisasi: Terapkan platform pemantauan model, jejak audit, dan kontrol akses.
- ✓ Laporkan kepada pimpinan eksekutif atau dewan direksi: Tata kelola harus memiliki visibilitas di tingkat kepemimpinan tertinggi.

Misalnya, sebuah bank global menjalankan "Kantor Kepatuhan Regulasi AI" internal yang melapor kepada CRO dan meninjau semua model di atas skor risiko yang telah ditentukan.

Kematangan Organisasi dan Tata Kelola AI

Seperti ukuran organisasi, tingkat kematangan suatu organisasi akan menentukan tingkat dan ketelitian formalitas tata kelola yang kemungkinan besar akan berhasil, menyeimbangkan kemampuan organisasi untuk mengelola dirinya sendiri dengan kemampuannya untuk mengidentifikasi dan mengelola risiko.

Organisasi yang lebih besar memiliki kematangan di pihak mereka tetapi dapat terhambat oleh birokrasi yang berlebihan. Organisasi yang lebih kecil memiliki keuntungan dalam hal kelincahan, tetapi mereka mungkin kesulitan untuk mengidentifikasi dan mengelola risiko secara efektif. Pernyataan-pernyataan sebelumnya mengasumsikan bahwa organisasi yang lebih besar memiliki kematangan yang lebih tinggi daripada yang lebih kecil, meskipun ini tidak selalu demikian. Beberapa organisasi yang lebih besar masih mempertahankan budaya

yang menyerupai budaya perusahaan rintisan, misalnya. Seperti topik-topik lain dalam bab ini, wawasan yang cerdas diperlukan untuk mengembangkan tingkat dan ketelitian tata kelola yang tepat yang memiliki potensi terbesar untuk efektivitas, yang mengarah pada keberhasilan proyek AI.

Fokus dari tingkat kematangan yang ada yang dibahas di sini akan paling sesuai dengan kematangan siklus pengembangan sistem (SDLC) suatu organisasi atau tata kelola TI atau teknologinya. Kedua model tata kelola ini tidak akan berbeda secara signifikan dari tata kelola AI.

Model Kematangan

Sebelum membahas detailnya, mari kita periksa terlebih dahulu model kematangan yang umum. Model kematangan adalah skala tingkat kematangan, masing-masing disertai dengan deskripsi singkat. Misalnya, model kematangan untuk COBIT memiliki enam tingkatan:

- **Tingkat O: Tidak Ada:** Tidak ada proses yang dapat dikenali sama sekali.
- **Tingkat 1: Awal/Ad Hoc:** Proses tidak terstruktur dan tidak dapat diprediksi. Keberhasilan bergantung pada upaya individu, bukan praktik yang dapat diulang.
- **Tingkat 2: Dapat Diulang:** Prosedur serupa mungkin ada, tetapi bersifat informal, tidak terdokumentasi, dan bervariasi menurut orang atau situasi.
- **Tingkat 3: Proses Terdefinisi:** Proses distandarisasi, didokumentasikan, dan dikomunikasikan. Orang-orang mengikuti pendekatan yang konsisten dan terdefinisi.
- **Tingkat 4: Terkelola dan Terukur:** Proses dipantau dan diukur. Manajemen dapat mengidentifikasi penyimpangan dan mengambil tindakan korektif.
- **Tingkat 5: Dioptimalkan:** Proses terus ditingkatkan berdasarkan data kinerja dan tujuan bisnis. Otomatisasi dan praktik terbaik dimanfaatkan.

Dalam konteks tata kelola AI, sebuah organisasi ingin menentukan tingkat kematangan minimum untuk proses yang dikelola melalui tata kelola AI. Sebuah organisasi dengan proses yang sudah ada dapat dinilai untuk menentukan tingkat kematangannya, dalam bentuk penilaian kesenjangan, untuk menentukan apakah proses tersebut setidaknya menunjukkan tingkat kematangan minimum atau apakah ada yang perlu ditingkatkan.

Catatan tentang Tingkat Kematangan

Tingkat kematangan rendah tidak boleh dianggap buruk, begitu pula tingkat kematangan tinggi tidak boleh dianggap baik. Terakhir, hanya sedikit organisasi yang perlu bercita-cita untuk mencapai tingkat kematangan 5. Sebaliknya, organisasi perlu menentukan tingkat kematangan yang sesuai, berdasarkan sektor industri, selera risiko, dan faktor lainnya. Selain itu, tidak semua proses perlu berada pada tingkat kematangan yang sama: sangat wajar jika satu proses berada pada tingkat 2 sementara proses lain berada pada tingkat 4.

Tingkat Kematangan Rendah

Ciri-ciri:

- Eksperimen dengan proyek percontohan

- Sedikit atau tidak ada keahlian internal
- Infrastruktur kebijakan minimal
- Kurangnya metodologi pengembangan formal

Strategi Tata Kelola:

- Fokus pada pendidikan dan kesadaran risiko
- Gunakan audit eksternal atau tinjauan etika untuk mengatasi kesenjangan internal
- Dokumentasikan keputusan meskipun informal (untuk penetapan dasar di masa mendatang)

Tingkat Kematangan Sedang

Ciri-ciri:

- Beberapa tim menerapkan model
- Standar informal mulai terbentuk
- Integrasi yang meningkat ke dalam proses bisnis
- Metodologi pengembangan dengan ketelitian dan kematangan sedang

Strategi Tata Kelola:

- Standardisasi terminologi dan dokumentasi model
- Buat repositori terpusat untuk kartu model dan penilaian
- Wajibkan proses verifikasi kasus penggunaan untuk sistem yang berhadapan dengan pelanggan

Tingkat Kematangan Tinggi

Ciri-ciri:

- Puluhan hingga ratusan model dalam produksi
- Model AI dengan latensi rendah dan waktu henti rendah
- Model AI yang menangani ribuan permintaan bersamaan
- Ketergantungan AI yang tinggi dalam sistem penghasil pendapatan
- Pengawasan peraturan dan paparan publik
- Metodologi pengembangan yang efektif dan matang

Strategi Tata Kelola:

- ✓ Sematkan gerbang tata kelola ke dalam AI Saluran operasional
- ✓ Lakukan audit pihak ketiga secara berkala
- ✓ Buat model risiko tingkat lanjut (misalnya, metrik keadilan, tolok ukur penjelasan)
- ✓ Gunakan infrastruktur pemantauan dan peringatan berkelanjutan

Imperatif Tata Kelola Spesifik Sektor Industri

Suatu struktur tata kelola AI yang efektif dan sesuai tidak dapat dibangun tanpa terlebih dahulu mempertimbangkan konteks industri. Hampir semua industri memiliki peraturan yang ada yang harus dipertimbangkan untuk membantu organisasi menghindari risiko kepatuhan, yang mengacu pada konsekuensi kegagalan untuk mematuhi hukum dan peraturan yang berlaku.

Industri yang Diatur

Contohnya termasuk keuangan, perawatan kesehatan, asuransi, energi, dan telekomunikasi.

Karakteristik Tata Kelola:

- Harapan hukum dan peraturan yang tinggi (misalnya, Basel III, HIPAA, GDPR)
- Persyaratan dokumentasi dan audit yang ekstensif
- Garis akuntabilitas yang jelas

Praktik Terbaik Tata Kelola:

- Gunakan tingkatan risiko formal untuk aplikasi AI, seperti yang diuraikan dalam Undang-Undang AI Uni Eropa
- Lakukan validasi model dan pengujian stres
- Buat keterlacakan dari sumber data ke prediksi
- Pertahankan log kontrol perubahan dan pembuatan versi

Misalnya, perusahaan asuransi kesehatan yang menerapkan AI untuk penolakan klaim harus memelihara log penjelasan dan menunjukkan keadilan klinis di seluruh kelompok demografis.

Teknologi yang Berorientasi pada Bisnis dan Konsumen

Contohnya meliputi e-commerce, media sosial, platform SaaS, aplikasi bisnis, perangkat keras TI, dan aplikasi seluler konsumen.

Karakteristik Tata Kelola:

- Siklus inovasi yang cepat
- Visibilitas publik yang tinggi
- Risiko yang muncul dari pengaruh dan kepercayaan algoritmik
- Kebijakan privasi yang mencakup penggunaan PII (Informasi Identitas Pribadi) oleh AI

Praktik Terbaik Tata Kelola:

- Prioritaskan mekanisme umpan balik pengguna
- Wajibkan fitur transparansi (misalnya, mengapa iklan/konten ini?)
- Lakukan penilaian dampak secara berkala pada sistem peringkat konten atau personalisasi

Industri dan Manufaktur

Contohnya termasuk logistik, otomotif, kimia, dan energi.

Karakteristik Tata Kelola:

- Risiko operasional (misalnya, keselamatan fisik, kerusakan lingkungan)
- Penggunaan tinggi pemeliharaan prediktif, robotika, dan algoritma optimasi

Praktik Terbaik Tata Kelola:

- Mitrakan AI dengan tim rekayasa keselamatan dan keandalan
- Pantau penyimpangan model yang dapat menyebabkan kegagalan peralatan
- Persyaratkan mekanisme pengaman dan pengesampingan manusia

Sektor Publik dan Nirlaba

Contohnya termasuk lembaga pemerintah, LSM, dan universitas.

Karakteristik Tata Kelola:

- Akuntabilitas yang lebih tinggi kepada publik
- Berorientasi pada misi (bukan berorientasi pada keuntungan)
- Lebih banyak batasan pada pengadaan dan kecepatan inovasi

Praktik Terbaik Tata Kelola:

- Publikasikan pernyataan dampak model terbuka
- Libatkan pemangku kepentingan komunitas dalam desain dan peninjauan
- Dokumentasikan dampak yang disengaja dan tidak disengaja pada populasi rentan

Produk dan Layanan: Tata Kelola Berdasarkan Dampak Kasus Penggunaan

Tidak semua sistem AI membawa risiko yang sama. Terlepas dari sektor industri, ukuran perusahaan, atau kematangan, tata kelola harus diskalakan secara proporsional dengan dampaknya.

Penggunaan AI Internal

Contohnya meliputi otomatisasi laporan pengeluaran, chatbot SDM, peramalan pendapatan, dan kampanye pemasaran.

- Risiko: Rendah
- Pendekatan tata kelola:
  - Gunakan dokumentasi yang disederhanakan
  - Tinjauan minimal kecuali jika data pribadi digunakan

AI Pendukung Keputusan

Contohnya meliputi segmentasi pelanggan, analitik prediktif, dan chatbot dukungan pelanggan.

- Risiko: Sedang
- Pendekatan tata kelola:
  - Validasi asumsi dan bias
  - Ungkapkan keterlibatan AI dalam dasbor atau laporan
  - Pantau akurasi dari waktu ke waktu

Pengambilan Keputusan Otomatis

Contohnya meliputi penyaringan pelamar kerja, aplikasi pinjaman, asisten diagnosis medis, dan aplikasi penyewa.

- Risiko: Tinggi
- Pendekatan Tata Kelola:
  - Pengawasan siklus hidup penuh
  - Audit bias dan pengujian keadilan
  - Proses penjelasan dan banding

Misalnya, sebuah perusahaan telekomunikasi menggunakan model AI untuk menilai kelayakan kredit pelanggan. Karena potensi dampak yang tidak proporsional, model tersebut harus menjalani tinjauan hukum dan etika sebelum diluncurkan dan masa percobaan pengawasan manusia yang ekstensif.

AI Berisiko Tinggi Waktu Nyata

Contohnya meliputi mengemudi otonom, ahli bedah robot AI, pencegahan penipuan, dan pengajaran K-12.

- Risiko: Sangat tinggi
- Pendekatan Tata Kelola:

- Simulasi tim merah dan pengujian lawan
- Pemantauan model waktu nyata dengan pemicu pematian otomatis
- Validasi manusia dalam lingkaran

Toleransi Risiko: Mengkalibrasi Tata Kelola Berdasarkan Selera Ketidakpastian

Toleransi risiko dan selera risiko suatu organisasi harus secara langsung memengaruhi intensitas tata kelolanya. Organisasi dengan selera dan toleransi risiko yang tinggi bersedia menghadapi lebih banyak ketidakpastian dibandingkan organisasi dengan selera dan toleransi risiko yang rendah.

Toleransi Risiko Rendah

Contohnya termasuk perbankan, jasa keuangan, perawatan kesehatan, dan energi.

Ciri-ciri Tata Kelola:

- Kuantifikasi risiko pra-implementasi
- Gerbang rilis konservatif
- Dokumentasi dan persetujuan komprehensif

Misalnya, perusahaan energi nuklir dengan toleransi rendah terhadap kegagalan model memerlukan persetujuan ganda dari manusia untuk perubahan operasional yang dipandu AI.

Toleransi Risiko Menengah

Contohnya termasuk perusahaan teknologi tinggi besar, produk konsumen, dan ritel.

Ciri-ciri Tata Kelola:

- Pengawasan bertingkat tergantung pada kasus penggunaan
- Otomatisasi parsial dengan pengawasan
- Komite etik terpusat untuk menyelesaikan ambiguitas

Toleransi Risiko Tinggi

Contohnya termasuk perusahaan rintisan, perusahaan media sosial, dan organisasi penelitian dan pengembangan.

Ciri-ciri Tata Kelola:

- Siklus iterasi lebih cepat dengan tinjauan pasca-hoc
- Kepemilikan risiko yang terdistribusi
- Pengamanan etika tetapi lebih sedikit batasan keras

Keselarasan Bisnis

Setiap program tata kelola, termasuk tata kelola AI, harus menunjukkan keselarasan bisnis. Keselarasan bisnis adalah konsep pengembangan kebijakan, proses, dan prosedur yang selaras dengan misi, tujuan, sasaran, ukuran, sektor, dan budaya organisasi. Semua topik yang telah dibahas sejauh ini (ukuran perusahaan, kematangan, sektor industri, konteks, dan toleransi risiko) adalah karakteristik dari keselarasan bisnis. Memahami keselarasan bisnis membantu menyesuaikan ruang lingkup tata kelola. Bagian ini mempertimbangkan karakteristik tambahan dari organisasi. Anda harus mempertimbangkan karakteristik ini sebagai generalisasi tetapi bukan sebagai satu-satunya cara untuk merancang program tata kelola.

Organisasi yang Didorong Inovasi

Contohnya termasuk perusahaan rintisan teknologi, perusahaan perangkat lunak, dan perusahaan energi alternatif.

Ciri-ciri:

- Siklus produk yang cepat
- Penekanan pada eksperimen
- Selera risiko yang relatif lebih tinggi

Pendekatan Tata Kelola:

- Jaga agar proses tetap lincah dan didorong oleh umpan balik
- Gunakan fase percontohan untuk menguji protokol tata kelola
- Tekankan pandangan ke depan yang etis daripada kepatuhan yang ketat

Organisasi yang Berfokus pada Reputasi

Contohnya termasuk penyedia layanan kesehatan, layanan keuangan, barang mewah, dan merek perhotelan.

Ciri-ciri:

- ✓ Paparan tinggi terhadap persepsi publik
- ✓ Beroperasi di domain yang sensitif terhadap kepercayaan

Pendekatan Tata Kelola:

- Membuat komitmen etika yang terbuka untuk publik
- Menggunakan dasbor transparansi
- Melibatkan PR/hukum dalam dewan peninjau

Organisasi yang berorientasi pada pengurangan biaya atau efisiensi

Contohnya termasuk jaringan restoran cepat saji, maskapai penerbangan, toko besar, dan layanan back-office.

Ciri-ciri:

- Penekanan pada otomatisasi dan AI operasional
- Kurang fokus pada fitur yang berorientasi pada pengguna

Pendekatan Tata Kelola:

- ➢ Mengevaluasi trade-off antara akurasi dan keadilan
- ➢ Fokus pada akuntabilitas dalam keputusan otomatis
- ➢ Mendokumentasikan ambang batas toleransi kesalahan dan mitigasi

2.5 PENGEMBANG, PENYEBAR, DAN PENGGUNA DALAM TATA KELOLA AI

Seiring AI diintegrasikan ke dalam proses bisnis inti, produk, dan layanan, tata kelola harus berkembang untuk mempertimbangkan seluruh komunitas pemangku kepentingan. Di dalam organisasi mana pun, tiga kelompok berbeda sering berinteraksi dengan sistem AI: pengembang, penyebar, dan pengguna. Peran-peran ini tumpang tindih di beberapa area tetapi berbeda secara signifikan dalam hal tanggung jawab, risiko, dan pengaruh terhadap perilaku dan hasil AI.

Memahami implikasi tata kelola dari setiap kelompok sangat penting untuk mengalokasikan akuntabilitas, merancang perlindungan yang efektif, dan memastikan bahwa

sistem AI beroperasi secara andal, adil, dan etis sepanjang siklus hidupnya. Bagian ini mengkaji perbedaan-perbedaan ini secara mendalam dan menyediakan kerangka kerja terstruktur untuk mengalokasikan tanggung jawab tata kelola di ketiga domain ini.

Definisi dan Batasan Peran

Pengembang AI

Pengembang AI adalah individu atau tim yang merancang, membangun, melatih, dan mengevaluasi model atau sistem AI. Pengembang AI meliputi:

- ❖ Ilmuwan data
- ❖ Insinyur pembelajaran mesin (ML)
- ❖ Insinyur data
- ❖ Ilmuwan riset
- ❖ Analis privasi data
- ❖ Analis keamanan siber

Area Fokus:

- Kurasi dan pra-pemrosesan data
- Pemilihan dan pelatihan model
- Evaluasi dan penyetelan kinerja
- Dokumentasi dan pengujian

Pihak yang Menerapkan AI

Pihak yang menerapkan AI adalah pemangku kepentingan yang bertanggung jawab untuk mengintegrasikan model AI ke dalam lingkungan produksi dan alur kerja operasional. Pihak yang menerapkan AI meliputi:

- Insinyur perangkat lunak
- Profesional operasi AI
- Manajer produk
- Tim infrastruktur

Area Fokus:

- ✓ Penerapan dan pemantauan model
- ✓ Infrastruktur dan skalabilitas
- ✓ Integrasi ke dalam aplikasi atau layanan
- ✓ Observabilitas pasca-penerapan

Pengguna AI

Pengguna AI adalah individu yang berinteraksi dengan sistem AI selama penggunaan operasionalnya. Pengguna AI meliputi:

- Pengguna internal (misalnya, staf SDM yang menggunakan alat penyaringan resume, atau pengembang yang mengarahkan AI LLM untuk menghasilkan atau memeriksa kode sumber)
- Pengguna eksternal (misalnya, pelanggan yang menerima rekomendasi yang dihasilkan AI)
- Pengambil keputusan yang mengandalkan wawasan AI (misalnya, petugas pinjaman, dokter, pemilik rumah)

Area Fokus:

- Interpretasi keluaran AI
- Ketergantungan operasional pada AI
- Pelaporan kesalahan, inkonsistensi, insiden, atau kerugian
- Penggunaan hak untuk mengajukan banding atau upaya pemulihan

Tanggung Jawab di Seluruh Siklus Hidup AI

Tata kelola AI harus mempertimbangkan berbagai titik kontak yang dimiliki setiap peran dalam siklus hidup AI. Dengan kata lain, ketiga peran yang dijelaskan sebelumnya memiliki tanggung jawab khusus dalam fase siklus hidup tata kelola AI.

Tabel 2.7 menggambarkan tanggung jawab pengembang AI pada umumnya. Misalnya, seorang ilmuwan data yang mengembangkan model penilaian kredit bertanggung jawab untuk menjalankan diagnostik keadilan dan mendokumentasikan keterbatasan yang diketahui, seperti penurunan kinerja pada populasi minoritas etnis.

Tabel 2.7
Tanggung Jawab Pengembang AI

Fase	Tugas Pengembang	Tanggung Jawab Tata Kelola
Pengumpulan data	Mengkurasi dan membersihkan kumpulan data	Periksa adanya bias representasional, pastikan persetujuan/privasi terjaga.
Desain model	Pilih algoritma, sesuaikan hyperparameter.	Dokumentasikan asumsi, pastikan dapat dijelaskan.
Evaluasi	Mengukur akurasi, keadilan, dan ketangguhan.	Lakukan pengujian bias, validasi terhadap tolok ukur, dan nilai keamanan.
Dokumentasi	Catat proses pengembangannya	Buat kartu model, lembar data, pengungkapan risiko.

Tabel 2.8 menggambarkan tanggung jawab seorang pengembang AI pada umumnya. Misalnya, seorang insinyur operasi AI bertanggung jawab untuk mengimplementasikan peringatan otomatis ketika prediksi model AI menyimpang dari akurasi yang diharapkan dan memastikan peringatan ini ditindaklanjuti dengan tepat.

Tabel 2.8
Tanggung Jawab Penyebar AI

Fase	Tugas Penyebar	Tanggung Jawab Tata Kelola
Integrasi	Sematkan model ke dalam aplikasi atau alat.	Pastikan UI/UX mendukung kemudahan penjelasan dan kontrol pengguna.
Penskalaan	Aktifkan akses dan manajemen beban	Terapkan kontrol akses dan pencatatan
Pemantauan	Lacak kinerja waktu nyata	Mendeteksi penyimpangan, memberikan peringatan jika terjadi anomali atau kegagalan.
Pemeliharaan	Kelola pembaruan model	Kontrol versi, prosedur pengembalian (rollback), validasi dampak

Tabel 2.9 menggambarkan tanggung jawab pengguna AI pada umumnya. Misalnya, seorang perekrut yang menggunakan filter resume berbasis AI harus memahami cara

mengesampingkan sistem ketika sistem tersebut mengecualikan kandidat yang memenuhi syarat berdasarkan kriteria yang sempit.

Tabel 2.9
Tanggung Jawab Pengguna AI

Fase	Tugas Pengguna	Tanggung Jawab Tata Kelola
Interaksi	Gunakan sistem AI dalam pengoperasiannya	Pahami peran AI dalam pengambilan keputusan, hindari kepercayaan buta.
Masukan	Mengidentifikasi masalah atau ketidaksesuaian	Laporkan anomali melalui saluran yang tepat
Pengawasan	Validasi keputusan yang dibantu AI.	Gunakan kebijaksanaan, mintalah peninjauan oleh manusia jika diperlukan.
Hak	Ajukan banding atas keputusan, upayakan transparansi.	Gunakan mekanisme ganti rugi jika terkena dampak.

Peluang dan Titik Pengungkit

Setiap peran tata kelola AI menghadirkan peluang penegakan yang unik, jika dimanfaatkan secara strategis.

Pengembang: Mengintegrasikan Tata Kelola ke dalam Desain

Pengembang memiliki kendali atas keputusan-keputusan penting yang membentuk perilaku sistem AI. Peluang tata kelola meliputi:

- **Mitigasi bias di sumbernya:** Pengembang mengontrol data pelatihan dan rekayasa fitur
- **Perangkat transparansi:** Membangun kemampuan menjelaskan ke dalam model (misalnya, menggunakan SHAP, LIME)
- **Pandangan etis:** Mempertimbangkan konsekuensi hilir selama eksperimen
- **Reproduksibilitas:** Menggunakan alur kerja, pembuatan versi, dan dokumentasi yang konsisten
- **Keamanan dan privasi:** Menilai aliran data dan merancang solusi sesuai dengan kebijakan privasi, kebijakan keamanan, dan persyaratan

Dorong pengembang untuk menggunakan daftar periksa desain etis dan berkontribusi pada lembar fakta model sebagai langkah standar dalam siklus pengembangan.

Penyebar: Mengoperasionalkan Tata Kelola dalam Skala Besar

Penyebar menerjemahkan niat pengembang menjadi dampak nyata di dunia nyata. Kontribusi tata kelola mereka meliputi:

- Mengintegrasikan infrastruktur pemantauan: Untuk mendeteksi penyimpangan kinerja dan keadilan.
- Mengotomatiskan pos pemeriksaan tata kelola: Untuk memastikan model tidak dapat diterapkan tanpa dokumentasi atau tinjauan risiko.
- Mengimplementasikan protokol pengembalian (rollback): Untuk mengatasi kerugian yang muncul atau kegagalan sistem.

Sebagai contoh, pengembang mengkonfigurasi lingkungan staging yang secara otomatis memeriksa dokumentasi model yang hilang sebelum mengizinkan promosi ke produksi.

Pengguna: Mendeteksi dan Melaporkan Masalah dari Garis Depan

Pengguna mengalami sistem AI dalam konteksnya, yang memungkinkan mereka untuk:

- Mendeteksi konsekuensi yang tidak diinginkan: Misalnya, rekomendasi yang membingungkan, perlakuan tidak adil, atau akurasi yang menurun.
- Memvalidasi output dengan pengetahuan domain: Misalnya, keahlian klinis untuk mendeteksi kesalahan diagnostik.
- Menginisiasi upaya hukum atau banding: Misalnya, memberi tahu tim tata kelola ketika dicurigai adanya kerugian.

Memberdayakan pengguna garis depan dengan pelatihan dan mekanisme pelaporan memungkinkan mereka menjadi pengawas risiko yang efektif. Pelatihan tata kelola AI, yang dibahas sebelumnya dalam bab ini, memberi tahu mereka apa yang perlu mereka ketahui, apa yang harus dicari, dan bagaimana melaporkan masalah.

Kebutuhan dan Harapan Sumber Daya

Pengembang, pengembang, dan pengguna AI membutuhkan pelatihan, alat, dan dukungan yang sesuai dengan kebutuhan unik setiap kelompok. Detailnya sebagai berikut.

Pengembang AI

Ingat, ini adalah individu atau tim yang merancang, membangun, melatih, dan mengevaluasi model atau sistem AI.

Kebutuhan:

- Alat untuk deteksi bias, penjelasan, dan pengujian ketahanan.
- Pedoman tata kelola yang disesuaikan dengan alur kerja pengembangan AI.
- Umpan balik dari pengembang dan pengguna untuk meningkatkan desain model.

Harapan:

- ✓ Kriteria keberhasilan yang jelas yang melampaui akurasi.
- ✓ Waktu yang dialokasikan untuk dokumentasi dan audit model.
- ✓ Ruang aman untuk menyampaikan kekhawatiran etis.

Pengembang AI

Ingat, ini adalah pemangku kepentingan yang bertanggung jawab untuk mengintegrasikan model AI ke dalam lingkungan produksi dan alur kerja operasional.

Kebutuhan:

- Akses ke model yang tervalidasi dan terdokumentasi dengan baik.
- Platform pemantauan yang menampilkan metrik yang relevan (penyimpangan, latensi, penggunaan).
- Panduan tentang integrasi pengawasan manusia dan antarmuka pengguna.

Harapan:

- Tanggung jawab bersama dengan pengembang untuk risiko pasca-implementasi.
- Sumber daya untuk mengelola pembaruan dan perubahan model secara bertanggung jawab.
- Partisipasi dalam rencana respons insiden.

Pengguna AI

Ingat, ini adalah individu yang berinteraksi dengan sistem AI selama penggunaan operasionalnya.

Kebutuhan:

- Pemahaman yang jelas tentang kapan dan bagaimana AI digunakan
- Pelatihan tentang cara menafsirkan hasil dan menggunakan kebijaksanaan
- Saluran pelaporan yang sederhana untuk kesalahan atau kerugian

Harapan:

- Hak untuk mendapatkan penjelasan, terutama untuk keputusan yang berdampak besar
- Mekanisme untuk banding, pembatalan, atau peninjauan oleh manusia
- Perlindungan dari pembalasan saat menyampaikan kekhawatiran

Konflik dan Ketidakselarasan Tata Kelola

Tanpa keselarasan, ketiga peran tata kelola AI dapat saling bertentangan.

Pengembang vs. Pengguna

- Ketegangan: Pengembang mengoptimalkan kinerja teknis; pengguna peduli pada kegunaan praktis dan kejelasan.
- Solusi tata kelola: Mengintegrasikan umpan balik pengguna ke dalam siklus pengembangan; mewajibkan pengujian kegunaan.

Pengembang vs. Penyebar

- Ketegangan: Pengembang mengharapkan serah terima yang lancar; penyebar sering berurusan dengan dokumentasi yang tidak lengkap atau kode yang tidak jelas.
- Solusi tata kelola: Membangun alur kerja operasi AI bersama dengan metadata dan ringkasan risiko yang diperlukan.

Pengguna vs. Penyebar

- Ketegangan: Pengguna mungkin menginginkan lebih banyak kendali dan kenyamanan daripada yang diizinkan oleh antarmuka, sementara penyebar dapat membatasi akses karena alasan keamanan atau skalabilitas.
- Solusi tata kelola: Melibatkan pengguna dalam pengujian UI/UX dan menyertakan opsi untuk peninjauan dan eskalasi oleh manusia.

Konflik lain dapat muncul dalam konteks pengembangan dan operasi sistem AI, seperti konflik antara TI dan keamanan siber, atau antara TI dan privasi. Konflik-konflik ini dan konflik lainnya dalam konteks pengembangan sistem tidak dibahas dalam buku ini.

Kontrol Tata Kelola Berbasis Peran

Menyesuaikan kontrol tata kelola untuk setiap peran memastikan pengembangan dan efektivitas operasional yang lebih besar.

Kontrol untuk Pengembang

- Kartu model: Meringkas tujuan, kinerja, dan keterbatasan model.
- Protokol pengujian keadilan: Menilai dampak yang berbeda selama pelatihan.
- Tinjauan kode sejawat untuk risiko: Tinjauan yang berfokus pada tata kelola di luar fungsionalitas.

- Tinjauan gerbang: Tinjauan formal memerlukan persetujuan manajemen sebelum melanjutkan ke langkah desain, pengembangan, dan pengujian selanjutnya.
- Persetujuan etika untuk kasus penggunaan baru: Tinjauan pra-penyebaran implikasi etika.

Kontrol untuk Penyebar

- Daftar periksa penyebaran: Mencakup dokumentasi, kepemilikan, dan jalur eskalasi.
- Log kontrol akses: Mencatat siapa yang dapat memicu atau memodifikasi model.
- Peringatan penyimpangan waktu nyata: Memberi tahu ketika perilaku model berubah secara signifikan.
- Tinjauan gerbang: Tinjauan formal memerlukan persetujuan manajemen sebelum melanjutkan ke langkah penyebaran selanjutnya.
- Jejak audit: Melestarikan riwayat perilaku sistem untuk investigasi.

Kontrol untuk Pengguna

- Pengungkapan transparansi: Tunjukkan kapan dan bagaimana AI terlibat dalam pengambilan keputusan.
- Mekanisme penggantian: Izinkan pengguna untuk membatalkan atau menantang keluaran AI.
- Saluran umpan balik: Izinkan pelaporan masalah langsung dari antarmuka.
- Prosedur penyelesaian sengketa: Sediakan proses banding formal atau peninjauan oleh manusia.

Koordinasi dan Integrasi Peran

Strukturisasi komunikasi lintas peran mendorong dialog antar peran. Misalnya:

- Retrospektif bersama setelah peluncuran sistem AI
- Komite etika lintas fungsi untuk kasus penggunaan berisiko tinggi
- Dasbor bersama yang menunjukkan kinerja model, metrik risiko, dan siklus umpan balik

Misalnya, setelah meluncurkan mesin rekomendasi, tim produk melakukan "Postmortem Tata Kelola" dengan ilmuwan data, agen layanan pelanggan, dan desainer UX untuk mengidentifikasi masalah kegunaan dan potensi risiko etika.

Organisasi mungkin ingin mempertimbangkan alat tata kelola AI terintegrasi yang mendukung dan melibatkan ketiga peran tersebut. Contohnya meliputi:

- Registri model dengan metadata audit
- Sistem peringatan yang mengarahkan insiden ke peran yang tepat
- Akses berbasis peran ke dokumentasi tata kelola

2.6 RINGKASAN

Menetapkan dan mengkomunikasikan harapan organisasi untuk tata kelola AI membutuhkan lebih dari sekadar dokumen kebijakan atau deklarasi eksekutif ini menuntut tindakan terstruktur berbasis peran dan keterlibatan lintas fungsi yang berkelanjutan. Organisasi dapat beralih dari prinsip-prinsip abstrak ke pelaksanaan tata kelola konkret dengan mendefinisikan tanggung jawab, mendorong kolaborasi, mengidentifikasi dan mendidik

pemangku kepentingan, menyesuaikan tata kelola dengan konteks, dan membedakan peran yang berinteraksi dengan sistem AI.

Landasan dari setiap program tata kelola terletak pada peran dan tanggung jawab yang didefinisikan dengan jelas. Organisasi harus memetakan akuntabilitas di seluruh kepemimpinan eksekutif, komite tata kelola AI, tim teknis, departemen hukum dan kepatuhan, keamanan siber dan privasi, unit bisnis, dan pengguna akhir. Dengan menyelaraskan tanggung jawab dengan fase spesifik siklus hidup AI dari pengumpulan data hingga penghentian model organisasi dapat meningkatkan pengawasan, mengurangi risiko, dan menumbuhkan wawasan etis. Alat akuntabilitas, seperti matriks RACI, peta peran siklus hidup, dan piagam tata kelola, membantu memastikan kejelasan dan mengurangi tumpang tindih atau kesenjangan dalam tanggung jawab.

Tata kelola AI yang efektif bukanlah domain dari satu fungsi saja, melainkan merupakan kerja tim. Kolaborasi lintas fungsi sangat penting untuk menangkap beragam keahlian dan perspektif. Tim hukum memahami nuansa peraturan, insinyur memahami arsitektur sistem, analis privasi memahami penggunaan data, dan petugas etika memfokuskan dampak manusia. Membentuk komite tata kelola lintas fungsi, dewan peninjau kasus penggunaan, dan proses dokumentasi kolaboratif memungkinkan kepemilikan dan pengambilan keputusan bersama.

Di luar struktur formal, memupuk budaya kolaboratif melalui pembelajaran sebaya, pemodelan kepemimpinan, dan keamanan psikologis memastikan bahwa tata kelola tidak hanya dipaksakan tetapi juga diinternalisasi. Membangun kapasitas organisasi melalui pelatihan dan kesadaran adalah pilar penting lainnya dari tata kelola AI yang sukses. Literasi AI harus meluas melampaui tim teknis untuk mencakup manajer bisnis, profesional hukum, dan staf operasional. Setiap kelompok membutuhkan pendidikan yang disesuaikan tentang terminologi, struktur tata kelola, tujuan strategis, dan kesadaran risiko. Program yang sukses menggabungkan pelatihan langsung, modul asinkron, alat bantu kerja, dan studi kasus dunia nyata. Strategi pelatihan yang matang tidak hanya memberikan pengetahuan tetapi juga memperkuat tata kelola sebagai praktik berkelanjutan yang terkait dengan kinerja. Kampanye, program penghargaan, dan tantangan yang digamifikasi dapat lebih menanamkan kesadaran ke dalam perilaku sehari-hari.

Tata kelola AI juga harus disesuaikan dengan konteks organisasi. Perusahaan rintisan dengan 20 karyawan, satu insinyur AI, dan kapasitas hukum yang terbatas tidak dapat diatur dengan cara yang sama seperti lembaga keuangan multinasional. Pendekatan tata kelola harus berskala dan fleksibel berdasarkan ukuran perusahaan, kematangan AI, industri, jenis produk, tujuan bisnis, budaya, dan toleransi risiko.

Sektor yang sangat diatur, seperti perawatan kesehatan atau keuangan, membutuhkan kontrol yang lebih ketat, sementara industri yang didorong oleh inovasi mungkin memprioritaskan pengawasan yang lincah dan ringan. Demikian pula, organisasi yang menggunakan AI untuk efisiensi internal mungkin mengadopsi tata kelola yang lebih terbatas daripada organisasi yang menerapkan AI dalam penggunaan yang berhadapan langsung dengan pelanggan dan berisiko tinggi. Tata kelola kontekstual memerlukan kerangka kerja

adaptif dan mekanisme pengawasan bertingkat untuk mencapai keseimbangan antara risiko dan responsivitas.

Tata kelola harus mempertimbangkan berbagai peran yang berinteraksi dengan sistem AI: pengembang, penyebar, dan pengguna. Pengembang mengontrol blok bangunan AI (data, algoritma, dan evaluasi) dan dengan demikian bertanggung jawab untuk menanamkan transparansi, keadilan, dan dokumentasi.

Penyebar bertanggung jawab untuk mengintegrasikan AI dengan aman ke dalam sistem, mempertahankan kinerja operasional, dan memastikan kapasitas pemantauan dan pengembalian. Seringkali diabaikan, pengguna memainkan peran penting dalam mengidentifikasi masalah yang muncul di lini depan, memvalidasi perilaku sistem dalam pengaturan dunia nyata, dan memulai eskalasi atau banding ketika terjadi kerugian. Tata kelola harus peka terhadap peran, dengan pelatihan, alat, dan harapan yang disesuaikan untuk setiap kelompok.

BAB 3
MEMPERBARUI KEBIJAKAN UNTUK AI

3.1 PENGAWASAN DI ERA PENGAMBILAN KEPUTUSAN OTONOM

Sebagian besar kebijakan yang ada di organisasi mencakup masalah manajemen sistem informasi umum, tetapi sistem AI memperkenalkan jenis situasi baru, seperti pengambilan keputusan otonom dan dampaknya terhadap privasi data.

Bagian ini mengkaji beberapa aspek kebijakan organisasi yang mungkin memerlukan revisi untuk memastikan mekanisme pengawasan dan akuntabilitas yang tepat di setiap fase siklus hidup AI. Meskipun organisasi sering berfokus pada kontrol tahap akhir, seperti pemantauan penyebaran atau pelaporan pasca-insiden, tata kelola yang efektif harus sadar akan siklus hidup. Ini berarti memperkenalkan kebijakan yang dapat ditegakkan dan ditinjau sejak saat kasus penggunaan diusulkan hingga saat model dihentikan.

Bagian ini menyajikan kerangka kebijakan siklus hidup yang dirancang untuk:

- Mengantisipasi dan mengurangi risiko sebelum terjadi kerugian
- Memungkinkan auditabilitas dan kepatuhan terhadap peraturan
- Memupuk kepercayaan publik pada sistem yang didukung AI

Seiring dengan terus berkembangnya kemampuan AI, standar tata kelola harus meningkat secara paralel. Masa depan AI yang dapat dipercaya tidak hanya bergantung pada apa yang kita bangun tetapi juga pada bagaimana kita mengaturnya di setiap langkah.

Model Kebijakan Siklus Hidup

Mengenai topik sistem AI yang memengaruhi kebijakan organisasi, mari kita mulai dengan gambaran besarnya. Setiap kali suatu organisasi berinvestasi dalam sistem informasi baru, sumber daya yang dikonsumsi dalam proses tersebut memerlukan pengamanan kebijakan dan titik pemeriksaan serta gerbang siklus hidup pengembangan. Ketika sistem informasi baru tersebut adalah sistem AI atau sistem yang diaktifkan AI, risiko dan manfaatnya meningkat, semakin memperkuat kebutuhan akan kebijakan untuk mencakup semua sudut pandang.

Mengapa Pengawasan Siklus Hidup Penting

Siklus hidup sistem AI pada dasarnya bersifat nonlinier dan iteratif. Asumsi yang salah yang dibuat selama akuisisi data dapat muncul sebagai perilaku model yang bias selama penerapan. Kurangnya dokumentasi selama pengembangan dapat menghambat audit atau tinjauan peraturan di masa mendatang.

Sebagai contoh, pada tahun 2020, sebuah perusahaan jasa keuangan besar menghadapi reaksi keras setelah sistem penilaian kredit berbasis AI-nya menawarkan batas kredit yang jauh lebih rendah kepada wanita daripada pria dengan profil keuangan yang sebanding. Investigasi mengungkapkan bahwa asumsi berbasis gender telah memengaruhi pemilihan fitur dan data pelatihan, namun tidak ada kebijakan yang mengharuskan tim untuk mendokumentasikan pilihan ini atau menilai implikasi etisnya. Pelajarannya jelas: Tata kelola

harus mengantisipasi risiko, bukan hanya bereaksi terhadapnya. Penemuan, analisis, dan keputusan yang dibuat saat mengidentifikasi risiko terkait AI harus diintegrasikan ke dalam prosedur "bisnis seperti biasa".

Fitur Inti dari Kebijakan yang Berorientasi pada Siklus Hidup

Kerangka kebijakan yang komprehensif harus:

- Mendefinisikan tanggung jawab dan jalur eskalasi untuk setiap fase siklus hidup dari awal hingga akhir.
- Dikalibrasi risiko (lebih ketat untuk sistem berisiko tinggi). Kasus penggunaan berisiko lebih tinggi (misalnya, dalam perawatan kesehatan, peradilan, dan penilaian pekerjaan) harus memicu proses peninjauan, dokumentasi, dan pengawasan yang lebih ketat.
- Memungkinkan auditabilitas dan pembuatan versi dari proses dan keputusan.
- Terintegrasi ke dalam alur kerja teknis, bukan paralel dengannya.
- Mencakup mekanisme untuk pemantauan berkelanjutan dan umpan balik, untuk melacak kinerja pasca-implementasi, pelaporan insiden, dan kalibrasi ulang model seiring evolusi konteks data dunia nyata.
- Menangani tata kelola data secara eksplisit, termasuk asal usul data, persetujuan, minimalisasi, kebijakan retensi, dan kepatuhan terhadap peraturan perlindungan data yang relevan (misalnya, GDPR, HIPAA).

Tidak cukup hanya membuat kebijakan—kebijakan tersebut harus dioperasionalkan melalui prosedur, gerbang, templat, daftar periksa, pencatatan, dan peran pengawasan yang dilatih untuk digunakan oleh tim.

Penilaian Kasus Penggunaan: Menyelaraskan Tujuan dan Risiko

Pengembangan kasus penggunaan berada di tahap awal pengembangan sistem AI. Kasus penggunaan adalah deskripsi singkat tentang konteks bisnis, beserta peran sistem AI dalam memfasilitasi atau meningkatkan proses bisnis dengan cara tertentu.

Haruskah AI Digunakan Sama Sekali?

Organisasi perlu mempertimbangkan dengan matang apakah solusi berbasis AI merupakan pendekatan yang tepat untuk setiap masalah bisnis atau teknis. Di beberapa organisasi, departemen TI (dan lainnya) begitu bersemangat untuk menggunakan alat baru mereka sehingga penilaian mereka mungkin terganggu, yang menyebabkan mereka menggunakan teknologi baru mereka untuk mengatasi setiap masalah, kendala, dan peluang yang mereka temui. Mengutip karakter Malcolm dalam film *Jurassic Park*, "... para ilmuwan Anda begitu sibuk memikirkan apakah mereka bisa melakukannya, sehingga mereka tidak berhenti berpikir apakah mereka harus melakukannya."

Sebelum satu baris kode pun ditulis, tim harus mengevaluasi apakah AI sesuai untuk masalah yang ada. Penilaian ini harus mempertimbangkan:

- Konteks sosial dari kasus penggunaan (misalnya, peradilan pidana, perawatan kesehatan, perhotelan).
- Apakah hasil yang diinginkan dapat memperkuat ketidakadilan atau menyebabkan kerugian.
- Ketersediaan dan kualitas data untuk mendukung pengembangan.

- Apakah sistem yang kurang kompleks atau lebih mudah diinterpretasikan dapat memenuhi tujuan yang sama. Sistem AI seringkali membutuhkan banyak sumber daya dan kompleks. Terkadang, solusi otomatisasi atau berbasis aturan yang sederhana sudah cukup. Tidak setiap masalah membutuhkan AI.

Beberapa organisasi memerlukan "memo justifikasi" untuk semua kasus penggunaan AI yang diusulkan, yang kemudian ditinjau oleh komite etika atau risiko AI. Memo ini mendokumentasikan tujuan, ruang lingkup, dampak pemangku kepentingan, dan keselarasan dengan nilai-nilai organisasi. Selain itu, sebelum pengembangan sistem AI dimulai, organisasi harus menggunakan kerangka kerja tata kelola lainnya, seperti Komite Pengarah TI, untuk mengikuti semua langkah yang diperlukan. Ini harus mencakup studi kelayakan dan penilaian risiko, untuk menentukan apakah proyek tersebut harus didanai dan, jika demikian, kapan harus dimulai dan sumber daya apa yang akan digunakan.

Kerangka Kerja Penilaian Risiko

Sistem AI yang diusulkan harus diberi tingkat klasifikasi atau peringkat risiko (seperti model klasifikasi yang terdapat dalam Undang-Undang AI Uni Eropa). Klasifikasi ini harus dilakukan sejak awal proyek pengembangan sistem AI baru, karena akan menentukan tingkat dan ketelitian pengembangan dan peninjauan desain, penilaian risiko, dokumentasi, dan aspek-aspek penting lainnya. Tanpa klasifikasi risiko formal, kasus penggunaan harus diberi peringkat risiko, seperti rendah, sedang, atau tinggi, berdasarkan kriteria seperti:

- ✓ Dampak pada individu atau kelompok yang dilindungi
- ✓ Paparan hukum atau peraturan
- ✓ Tingkat otomatisasi dalam pengambilan keputusan

Kasus penggunaan berisiko tinggi mungkin memerlukan audit pihak ketiga, masukan pemangku kepentingan eksternal, atau persetujuan eksekutif sebelum dilanjutkan.

Manajemen Risiko yang Disesuaikan untuk AI

Kerangka kerja manajemen risiko suatu organisasi kemungkinan perlu disesuaikan untuk mengakomodasi penggunaan sistem AI. Salah satu aktivitas kunci adalah penetapan tingkat risiko atau klasifikasi untuk setiap sistem AI prospektif sebelum desain dan pengembangannya. Kerangka kerja, seperti Undang-Undang AI Uni Eropa, mendefinisikan tingkat risiko dengan aktivitas yang ditentukan untuk dilakukan selama pengembangan dan operasi berkelanjutan.

Tata Kelola TI dan AI Terpisah atau Tergabung?

Organisasi dengan tata kelola TI yang efektif dapat memilih untuk menambahkan langkah-langkah khusus AI ke dalam proses tata kelola TI dan siklus hidup pengembangan sistem mereka. Menggabungkan tata kelola AI dan TI ke dalam satu proses yang mengakomodasi keduanya seringkali lebih efektif daripada mempertahankan proses yang terpisah.

Sistem AI menghadirkan risiko baru yang tidak sepenuhnya ditangani oleh manajemen risiko perangkat lunak tradisional:

- **Ketidakpastian epistemik:** Sistem AI mungkin secara yakin salah karena adanya kesenjangan data atau pelatihan yang tidak tepat.
- **Bias otomatisasi:** Operator manusia cenderung mempercayai keluaran model meskipun terdapat kesalahan.
- **Perilaku yang muncul secara tiba-tiba:** Hasil yang tidak terduga atau tidak direncanakan yang melampaui tujuan sistem AI, terutama dalam sistem multi-agen atau generatif.

Di era AI, integritas data adalah sesuatu yang harus kita fokuskan jauh lebih banyak daripada di masa lalu. Kerangka kebijakan harus memastikan bahwa identifikasi risiko terjadi secara terus-menerus, bukan hanya selama perencanaan dan pengembangan.

Sebagai contoh, sistem rekomendasi AI ritel ditemukan memperkuat stereotip gender dengan menampilkan produk rumah tangga kepada wanita dan elektronik kepada pria. Meskipun secara teknis berkinerja baik, hal itu melanggengkan bias budaya, mengungkapkan titik buta dalam identifikasi risiko.

Kebijakan harus mencakup tinjauan risiko terjadwal dan berbasis pemicu, seperti:

- Setelah pembaruan data atau model utama
- Ketika indikator pemantauan bergeser atau kinerja menurun
- Sebagai tanggapan terhadap umpan balik publik atau perubahan hukum

Tinjauan harus didokumentasikan, dengan pemilik risiko yang jelas ditugaskan. Tindak lanjut harus diperlukan untuk mengatasi semua risiko yang teridentifikasi.

Etika Berdasarkan Desain: Dari Prinsip ke Praktik

Subjek data seringkali memiliki gagasan yang cukup jelas tentang benar dan salah, termasuk akuisisi dan penggunaan data tentang mereka oleh suatu organisasi. Organisasi yang mempertimbangkan penggunaan sistem AI harus mengembangkan pedoman etika formal yang menjadi bagian dari siklus pengembangan sistem AI standar.

Mengoperasionalkan Etika AI

Nilai-nilai abstrak, seperti keadilan dan martabat, harus diterjemahkan ke dalam perlindungan praktis dan langkah-langkah yang dapat ditindaklanjuti. Ini termasuk:

- Penilaian dampak etis yang terkait dengan kasus penggunaan setiap model
- Mekanisme untuk mengungkap "zona keruntuhan moral," di mana tanggung jawab secara tidak adil dialihkan kepada individu dengan otoritas rendah (misalnya, agen layanan pelanggan yang menjelaskan keputusan AI yang tidak dapat mereka kendalikan)
- Keterlibatan pemangku kepentingan (terutama komunitas yang terdampak) dalam pilihan desain

Seperti analisis risiko atau analisis dampak privasi, analisis etika dari sistem AI yang diusulkan harus konsisten, dapat diulang, dan dilakukan oleh personel yang terlatih. Penilaian tersebut dapat terdiri dari beberapa pertanyaan etika sederhana, seperti:

- Bisakah sistem ini memperburuk ketidakadilan sosial?
- Apakah populasi rentan terdampak secara tidak proporsional?
- Bisakah pengguna menentang atau mengajukan banding atas keputusan tersebut?

Pendekatan yang lebih formal dapat berupa kuesioner yang lebih panjang atau bahkan alat daring untuk organisasi yang melakukan banyak penilaian semacam itu.

Membangun Etika ke dalam Desain Teknis

Perancang dan pengembang sistem AI harus dilengkapi dengan alat-alat seperti:

- Perangkat bantu keadilan (misalnya, Aequitas, AI Fairness 360)
- Pustaka penjelasan (misalnya, SHAP, LIME)
- Dasbor deteksi bias selama validasi model

Kebijakan harus mewajibkan penggunaan alat-alat tersebut dan mengharuskan output didokumentasikan dalam catatan bisnis.

Akuisisi dan Penggunaan Data

Sistem AI mengonsumsi sejumlah besar data selama pelatihan dan penggunaan operasional. Organisasi yang mempertimbangkan penggunaan sistem AI yang akan memproses data mereka harus sudah memiliki tata kelola data yang efektif. Jika tidak, pengembangan kebijakan, prosedur, dan alat tata kelola data harus dianggap sebagai prasyarat prioritas tinggi untuk AI.

Silsilah dan Dokumentasi Data

Setiap proyek AI harus mencakup pengembangan dan pemeliharaan dokumen asal usul data (juga dikenal sebagai silsilah data), yang mencakup:

- Sumber setiap dataset
- Persyaratan lisensi atau penggunaan
- Keterbatasan yang diketahui (misalnya, kurangnya representasi, ketidakseimbangan kelas)

Misalnya, AI SDM yang dilatih hanya berdasarkan data perekrutan masa lalu mungkin mengandung bias atau diskriminasi historis. Tinjauan silsilah akan mengungkap hal ini dan memicu langkah mitigasi yang diamanatkan kebijakan, seperti penyeimbangan ulang atau pengecualian fitur yang bias.

Penelusuran Silsilah Data Seharusnya Sudah Ada

Penelusuran silsilah data seharusnya tidak hanya diperlukan untuk proyek AI, tetapi untuk setiap proyek sistem TI, baik internal, eksternal, atau daring, seperti SaaS.

Persetujuan, Privasi, dan Data Sintetis

Kebijakan manajemen data dan/atau privasi suatu organisasi juga harus:

- ✓ Mensyaratkan persetujuan yang diinformasikan jika sesuai dan diwajibkan oleh hukum
- ✓ Menangani penggunaan perantara data pihak ketiga
- ✓ Mewajibkan teknik yang menjaga privasi (misalnya, privasi diferensial) di domain sensitif

Organisasi yang menggunakan data sintetis juga harus menetapkan perlindungan kebijakan untuk memastikan realisme tanpa mengorbankan kerahasiaan data pribadi.

Pengembangan Model: Batasan untuk Penciptaan

Organisasi perlu mengembangkan pendekatan standar untuk pengembangan model dan sistem AI, sehingga sebagian besar atau semua sistem tersebut memiliki konsistensi yang membuat dukungan terhadapnya lebih mudah. Konsep ini mirip dengan adopsi bahasa pemrograman standar untuk menghindari jumlah bahasa unik yang berlebihan yang dipilih sesuka hati pengembang.

Menegakkan Reproduksibilitas dan Akuntabilitas

Pengembangan model AI seharusnya bukan proses yang bersifat manual, melainkan hasil dari disiplin rekayasa. Selain persyaratan untuk siklus pengembangan sistem (SDLC), kebijakan harus mencakup:

- Daftar singkat alat, sistem, dan platform AI yang disetujui, serupa dengan standar bahasa pemrograman SDLC
- Sistem kontrol versi (misalnya, Subversion, DVC, MLflow) dengan aturan check-in dan check-out yang terkelola
- Permintaan perubahan model yang ditinjau oleh rekan sejawat atau disetujui oleh manajer
- Kartu model yang menjelaskan tujuan, keterbatasan, dan hasil pengujian
- Ketelusuran sumber data pelatihan dan pipeline pra-pemrosesan
- Pengujian otomatis untuk kinerja, bias, dan ketahanan sebagai bagian dari pipeline CI/CD
- Antarmuka dan protokol komunikasi yang disetujui (misalnya, Protokol Konteks Model—MCP) untuk menstandarisasi interaksi model dengan API, layanan, atau agen

Beberapa perusahaan menerapkan "aturan dua orang," di mana tidak ada model yang dapat maju ke pengujian kecuali telah ditinjau dan disetujui oleh praktisi AI yang berkualifikasi lainnya. Tetapi pendekatan yang lebih formal adalah proses gerbang dengan tinjauan dan persetujuan formal.

Arsitektur Model yang Aman dan Dapat Diinterpretasikan

Kebijakan pengembangan AI suatu organisasi harus mengarahkan tim untuk menghindari model yang tidak perlu buram, terutama dalam kasus penggunaan berisiko tinggi. Hal ini mencakup:

- Menegakkan penggunaan model yang dapat diinterpretasikan kecuali jika pertimbangan kinerja didokumentasikan dan disetujui.
- Membutuhkan justifikasi dan persetujuan formal untuk arsitektur yang kompleks, seperti jaringan saraf dalam.

Misalnya, sistem berbasis aturan mungkin lebih tepat untuk keputusan kelayakan daripada model ensemble kotak hitam.

Pelatihan dan Pengujian: Mempersiapkan Diri untuk Dunia Nyata

Keberhasilan platform AI sangat bergantung pada keberhasilan pelatihan dan pengujian yang direncanakan dengan baik. Tanpa pelatihan yang efektif, sistem AI kemungkinan besar tidak akan memberikan hasil yang diharapkan.

Karena buku ini berfokus pada tata kelola AI daripada rekayasa AI, buku ini tidak membahas detail tentang apa yang membentuk pelatihan dan pengujian yang efektif, sebagian karena terlalu banyak variabel yang berperan. Sebaliknya, sebagai seorang profesional tata kelola, Anda atau seseorang di organisasi Anda perlu menyusun aturan bisnis dalam tata kelola AI untuk mendefinisikan persyaratan untuk pelatihan dan pengujian model. Namun, saya memberikan kategori dan prinsip pelatihan dan pengujian; Pertama, tentang pelatihan model:

- Data pelatihan akurat dan lengkap
- Hak privasi subjek data dijaga (masa lalu, sekarang, dan masa depan)

Perhatikan bahwa diskusi ini berfokus pada pelatihan model dan sistem AI, bukan pelatihan individu, yang dibahas di Bab 2.

Berikut adalah kategori pengujian:

- Akurasi
- Keadilan
- Bias
- Kemampuan menjelaskan
- Kinerja
- Keamanan

Standar Validasi yang Ketat

Kebijakan untuk pelatihan dan pengujian harus dikembangkan untuk memastikan bahwa model dan sistem AI berkinerja sesuai kebutuhan dan juga adil, andal, dan tangguh.

Kebijakan harus mendefinisikan:

- Tolok ukur kinerja yang disesuaikan dengan tujuan kasus penggunaan
- Analisis subkelompok untuk mendeteksi dampak yang berbeda, seperti bias berbasis gender atau ras
- Kontrol untuk kebocoran data dan overfitting

Semua pengujian, termasuk rencana pengujian, hasil pengujian, dan persetujuan pengesampingan apa pun (misalnya, persetujuan manajemen untuk peluncuran jika kinerja sistem tidak memenuhi persyaratan) harus didokumentasikan sebagai bagian dari catatan bisnis permanen sistem.

Sebagai contoh, model deteksi penipuan tidak hanya harus menjaga akurasi tetapi juga meminimalkan false positive untuk pengguna yang sah. Kebijakan mungkin memerlukan pengujian skenario dengan alur transaksi dunia nyata, serta tingkat toleransi terhadap false positive.

Pengujian Red Team dan Pengujian Adversarial

Seperti sistem informasi baru lainnya (AI atau bukan), sistem AI baru harus menjalani pengujian keamanan untuk menilai ketahanannya terhadap berbagai jenis serangan. Jenis pengujian yang harus dilakukan meliputi:

- Pengujian input adversarial
- Skenario injeksi prompt
- Simulasi penyalahgunaan pengguna
- Serangan siber konvensional

Kebijakan dapat menetapkan peran "tim merah" keamanan siber yang bertugas melakukan pengujian stres pada sistem dan mencatat hasilnya untuk tinjauan tata kelola.

Hasil Pengujian Merupakan Gerbang Vital

Seperti yang dibahas di bagian ini, rencana pengujian dan hasil pengujian harus didokumentasikan dan diarsipkan. Tata kelola AI harus mencakup titik pemeriksaan di akhir pengujian, sehingga manajemen dapat memahami hasil pengujian, mengajukan pertanyaan tentangnya, dan membuat keputusan untuk melanjutkan atau tidak melanjutkan proyek pengembangan sistem AI.

Penyebaran dan Pemantauan: Menjaga Gerbang

Ketika kita sampai pada titik penyebaran, kebijakan tata kelola AI harus mensyaratkan setidaknya satu kesempatan bagi manajemen untuk meninjau semua hasil analisis dan pengujian sebelum manajemen memberikan persetujuan formal agar sistem AI dipindahkan ke produksi.

Kontrol Pra-Penyebaran

Sebelum penyebaran, kebijakan harus mensyaratkan:

- ✓ Titik pemeriksaan tinjauan etika dan kepatuhan
- ✓ Tinjauan penilaian dan mitigasi risiko
- ✓ Paket dokumentasi lengkap yang diajukan
- ✓ Rencana pemantauan yang disetujui oleh badan tata kelola yang relevan

Dalam model risiko multi-tingkat, tingkat ketelitian untuk langkah-langkah ini dapat bervariasi tergantung pada tingkat risiko dan konteksnya.

Sebagai contoh, sistem AI perawatan kesehatan tidak hanya harus menunjukkan efektivitas model tetapi juga kesiapan komunikasi pasien (misalnya, protokol keterlibatan manusia dan privasi).

Pemantauan Pasca-implementasi

Setelah sistem AI memasuki layanan produksi, kebijakan tata kelola AI harus mensyaratkan:

- ➢ Pelacakan kinerja secara real-time terhadap baseline
- ➢ Peringatan untuk penyimpangan model, data di luar distribusi, dan metrik keadilan
- ➢ Pengawasan manusia, terutama untuk keputusan yang berdampak besar

Dasbor pemantauan harus dapat diakses oleh pemangku kepentingan non-teknis, termasuk petugas kepatuhan dan mungkin regulator. Karena audiens yang beragam ini, pemantauan harus mencakup metrik dan pelaporan terkait bisnis, bukan hanya berfokus pada kinerja teknis.

Organisasi dengan Kematangan Proses Rendah: Waspadalah

Organisasi dengan kematangan proses yang rendah, khususnya dalam pengembangan sistem dan manajemen layanan TI, akan menghadapi risiko masalah yang jauh lebih tinggi dengan sistem AI mereka. Jika ini terjadi di organisasi Anda, saya mendorong Anda untuk mengidentifikasi seseorang yang dapat membahas masalah ini secara terbuka dan jelas dengan manajemen senior.

Dokumentasi dan Pelaporan: Membangun Memori Institusional

Anda telah melihatnya disebutkan di seluruh bagian ini, dan berikut ini disebutkan lagi. Setiap program tata kelola teknis, termasuk untuk AI, harus mencakup dokumentasi yang menyeluruh. Dokumentasi ini membantu mereka yang bekerja pada sistem AI di kemudian hari untuk memahami bagaimana sistem tersebut awalnya dirancang, diuji, dibangun, dan dioperasikan. Anda dapat melakukan rekayasa balik beberapa aspek sistem tetapi bukan persyaratannya, hasil pengujian, dan keputusan bisnis; hanya serangkaian catatan bisnis yang dapat memenuhi kebutuhan tersebut.

Dokumentasi yang Dinamis

Manajemen dan pemeliharaan sistem AI harus didokumentasikan sehingga semua pihak yang terlibat lebih memahami desainnya, keterbatasan desain, parameter operasi, harapan kinerja, dan banyak lagi. Sistem AI harus dapat diaudit. Kebijakan harus mensyaratkan:

- Dokumentasi model diperbarui setelah setiap perubahan signifikan.
- Log keputusan untuk poin tinjauan etika dan hukum.
- "Snapshot penjelasan" untuk model yang diterapkan.

Selain itu, sistem AI harus dapat diaudit. Sekali lagi, ini berarti bahwa tujuan sistem, persyaratan, desain, pelatihan, pengujian, dan operasinya harus didokumentasikan, seperti sistem informasi lainnya.

Pelaporan Internal dan Eksternal

Tergantung pada yurisdiksi, risiko sistem, dan konteks, kebijakan juga harus mewajibkan:

- Pelaporan internal reguler kepada komite tata kelola
- Laporan transparansi eksternal (dianonimkan sesuai kebutuhan)
- Pengungkapan kepada pengguna atau publik, terutama untuk sistem yang memengaruhi hak

Manajemen Insiden: Perencanaan untuk Kegagalan

Respons insiden keamanan (IR) bukanlah hal baru—ini telah menjadi andalan praktik standar sejak tahun 1990-an, dan bahkan lebih awal untuk beberapa organisasi. Seperti disiplin ilmu lain dalam TI, seperti keamanan siber dan privasi, prosedur respons insiden yang formal dan terdokumentasi perlu diperluas ke semua sistem AI yang digunakan.

Organisasi dengan program respons insiden keamanan yang berfungsi akan menemukan bahwa peningkatan pada program tersebut akan mewakili peningkatan perubahan yang relatif kecil.

Mendefinisikan dan Meningkatkan Insiden

Tidak semua kesalahan sistem AI adalah insiden. Kebijakan tata kelola AI dan respons insiden keamanan harus mengklarifikasi:

- Apa yang dianggap sebagai insiden AI (misalnya, penggunaan tanpa izin, keputusan yang bias, halusinasi sistem, kebocoran data, penyimpangan)
- Batas waktu dan ambang batas pelaporan

- Templat komunikasi untuk regulator, pelanggan, atau publik

Satu atau lebih perubahan kebijakan ini akan memerlukan peningkatan atau penambahan pemantauan. Lagipula, sulit untuk menyatakan bahwa suatu insiden sedang terjadi (atau telah terjadi) jika tidak ada kemampuan deteksi yang tersedia. Oleh karena itu, mungkin perlu untuk menggunakan atau mengembangkan kemampuan pemantauan untuk memungkinkan deteksi ini. Misalnya, jika chatbot mulai menghasilkan respons yang tidak pantas, kebijakan mungkin mengharuskan penonaktifan model, pemberitahuan kepada pengguna yang terpengaruh, dan pemicuan analisis akar penyebab yang lengkap.

Diperlukan Buku Panduan AI

Dalam respons insiden keamanan, praktik standar adalah mendokumentasikan rencana respons insiden keamanan yang luas dan menyeluruh yang mencakup peran dan tanggung jawab, tingkat keparahan, pelaporan manajemen, dan pemberitahuan kepada pihak internal dan eksternal. Praktik standar juga mencakup pengembangan sejumlah "buku panduan," yang merupakan instruksi langkah demi langkah yang terperinci untuk diikuti ketika jenis insiden tertentu terjadi.

Satu atau lebih buku panduan kemungkinan harus ditulis untuk mencakup jenis insiden yang unik untuk sistem AI. Seperti buku panduan IR lainnya, buku panduan baru yang terkait dengan sistem AI harus diuji, dan personel harus dilatih.

Akar Penyebab dan Perbaikan

Kebijakan dan prosedur respons insiden standar seringkali memerlukan beberapa tingkat tinjauan pasca-insiden. Tujuan dari tinjauan tersebut adalah untuk menilai seberapa efektif tindakan deteksi dan respons organisasi dieksekusi, dan apakah ada penyesuaian terhadap kebijakan, desain sistem, atau prosedur respons insiden yang diperlukan. Perubahan tersebut membantu organisasi menghindari insiden serupa di masa mendatang dan/atau meresponsnya dengan lebih efektif.

Setelah insiden:

- Tim multidisiplin melakukan postmortem.
- Kesenjangan tata kelola diidentifikasi.
- Kebijakan dan pelatihan diperbarui sesuai dengan temuan.

Kebijakan dan budaya organisasi harus menekankan retrospektif tanpa menyalahkan yang mendorong pengungkapan masalah sistemik, bukan menjadikan individu sebagai kambing hitam.

3.2 EVALUASI & PERBARUI KEBIJAKAN PRIVASI-KEAMANAN DATA AI

Kerangka kerja manajemen data, privasi, dan keamanan tradisional tidak dirancang dengan mempertimbangkan AI. Kerangka kerja tersebut berfokus pada pengendalian akses, menjaga kerahasiaan, dan mencegah pelanggaran, penting tetapi tidak cukup dalam konteks sistem AI yang belajar, menggeneralisasi, dan menyimpulkan dari data. Dengan AI, data bukan hanya aset statis, tetapi juga bahan bakar untuk model pengambilan keputusan, seringkali dalam lingkungan yang berisiko tinggi.

Oleh karena itu, organisasi harus mengevaluasi dan merevisi kebijakan manajemen data, privasi, dan keamanan yang ada untuk mencerminkan tantangan khusus AI. Ini termasuk mengatasi kemampuan AI untuk menyimpulkan informasi sensitif, risiko identifikasi ulang dalam kumpulan data anonim dan pseudonim, dan implikasi data sintetis dan augmentasi data. Bagian ini menawarkan pendekatan terstruktur untuk menilai dan memperbarui kebijakan privasi dan keamanan agar siap untuk AI.

Mengevaluasi dan memperbarui kebijakan privasi dan keamanan data yang ada bukan lagi pilihan ini sangat penting untuk:

- Melindungi hak individu dalam menghadapi sistem AI yang canggih
- Memenuhi kewajiban hukum berdasarkan regulasi AI yang terus berkembang
- Memastikan kepercayaan subjek data dan pemangku kepentingan terhadap proses yang digerakkan oleh AI

Organisasi yang bertindak sekarang untuk mempersiapkan kerangka kerja tata kelola mereka di masa depan akan berada pada posisi terbaik untuk berinovasi secara bertanggung jawab di era sistem cerdas.

Mengapa AI Mengganggu Tata Kelola Data Tradisional

Sistem AI menggunakan data dengan cara yang jarang ditemukan sebelumnya. Oleh karena itu, organisasi yang menerapkan tata kelola AI perlu meninjau dan merevisi tata kelola data mereka.

AI Sangat Membutuhkan Data

Model AI dan ML, terutama LLM dan jaringan saraf dalam, membutuhkan kumpulan data yang luas dan beragam. Kebutuhan akan data ini menciptakan insentif untuk:

- ✓ Menggabungkan data dari berbagai sumber
- ✓ Membeli atau mengambil data dari pihak ketiga
- ✓ Menggunakan kembali data yang dikumpulkan untuk tujuan yang tidak terkait

Kebijakan privasi tradisional seringkali gagal memperhitungkan perilaku khusus AI ini dan mungkin kurang memiliki kontrol untuk:

- ➢ Pembatasan tujuan dalam sistem AI
- ➢ Akses berkelanjutan ke data pasca-pelatihan
- ➢ Tata kelola data pelatihan pihak ketiga
- ➢ Praktik penyimpanan data

Risiko Inferensi dan Reidentifikasi

Model AI dapat menyimpulkan sifat sensitif, seperti ras, pandangan politik, atau orientasi seksual, bahkan jika label tersebut tidak secara eksplisit ada. Selain itu, model yang dilatih pada data yang seharusnya dianonimkan atau dipseudonimkan telah terbukti membocorkan informasi identitas, terutama ketika terpapar input eksternal atau dipicu dengan cara yang merugikan.

Sebagai contoh, pada tahun 2021, para peneliti menunjukkan bahwa model bahasa bergaya GPT dapat menghafal dan mengulang data sensitif (misalnya, nomor Jaminan Sosial, email) dari korpus pelatihan mereka, bahkan jika data tersebut telah sedikit disamarkan. Kelemahan ini menantang asumsi kontrol privasi tradisional.

Kesenjangan Kebijakan Privasi dan Ancaman Khusus AI

AI mengubah cara penggunaan PII (Informasi Identitas Pribadi), dan keterbatasan AI saat ini menimbulkan tantangan terhadap "hak untuk dilupakan" dan kebutuhan privasi lainnya. Tidak ada bidang lain di bidang TI yang begitu berpengaruh seperti privasi. Tata kelola privasi merupakan prasyarat untuk pengembangan tata kelola AI. Tanpa tata kelola privasi, organisasi lebih cenderung melanggar hukum privasi atau harapan masyarakat.

Kekurangan Kebijakan Umum

Sebagian besar kebijakan privasi dan keamanan data yang ada di organisasi:

- Tidak menentukan berapa lama data pelatihan disimpan setelah pengembangan model
- Kurangnya kontrol untuk serangan inversi model atau inferensi keanggotaan
- Menganggap minimalisasi data hanya berlaku untuk penyimpanan, bukan untuk apa yang disimpan oleh model
- Tidak menyertakan hak subjek data dalam keluaran model (misalnya, hak untuk penjelasan, penolakan, penghapusan)

Akibatnya, risiko privasi dan keamanan dapat menyebar secara diam-diam dari data ke model hingga prediksi.

Ancaman Privasi Khusus AI

Ancaman privasi khusus AI yang perlu diperhatikan meliputi:

- **Inversi model:** Merekonstruksi data masukan dengan menyelidiki model.
- **Inferensi keanggotaan:** Upaya untuk menentukan apakah suatu catatan tertentu merupakan bagian dari set pelatihan sistem AI.
- **Peracunan model:** Penyerang dengan sengaja menyuntikkan data palsu ke dalam alur pelatihan.
- **Perluasan fungsi:** Memperluas penggunaan data di luar tujuan awal yang dinyatakan melalui penggunaan kembali AI.
- **Hafalan yang tidak disengaja:** LLM dapat menyimpan dan mereproduksi fragmen data pelatihan sensitif, bahkan tanpa perintah langsung.

Ancaman-ancaman ini menuntut sikap kebijakan baru: sikap yang melihat privasi bukan sebagai keadaan sekali waktu tetapi sebagai pertimbangan sistemik yang berkelanjutan di seluruh data dan model.

Area Kebijakan Utama yang Perlu Dievaluasi dan Diperbarui

Bagian ini membahas beberapa area di mana organisasi mungkin perlu memperbarui kebijakan privasi mereka.

Batasan Tujuan dalam Pipeline Dinamis

Pengembangan AI sering kali melibatkan penggunaan kembali data di berbagai eksperimen, penyempurnaan, dan pemodelan ensemble. Kebijakan privasi (termasuk, dalam beberapa kasus, pemberitahuan praktik privasi yang dapat diakses publik—NOPP) harus:

- Menentukan kondisi penggunaan kembali yang diizinkan
- Memerlukan persetujuan baru jika kasus penggunaan asli berubah secara signifikan
- Menetapkan pencatatan dan jejak audit untuk penggunaan setiap dataset di berbagai versi model

Standar Anonimisasi dan Pseudonimisasi yang Ditingkatkan

Teknik anonimisasi dan pseudonimisasi tradisional mungkin tidak efektif terhadap serangan reidentifikasi AI. Persyaratan kebijakan yang diperbarui harus mencakup:

- Metrik risiko yang terukur (misalnya, k-anonimitas, l-diversitas)
- Privasi diferensial untuk domain sensitif
- Anonimisasi ulang wajib saat menggabungkan kumpulan data

Memperkuat Anonymisasi

Pertimbangkan pembuatan data sintetis yang dikombinasikan dengan privasi diferensial untuk kumpulan data publik dan secara rutin audit perilaku model untuk pola memorisasi.

Kontrol Retensi Data Berbasis Siklus Hidup

Organisasi perlu mengubah kebijakan privasi mereka untuk membedakan antara:

- Data input mentah
- Data yang ditransformasikan/berbasis fitur
- Kumpulan data berlabel
- Artefak model

Setiap lapisan memerlukan kebijakan retensi dan penghapusan yang berbeda. Pertimbangan harus mencakup:

- ✓ Apakah data pelatihan harus dihapus setelah konvergensi model
- ✓ Apakah versi data sintetis atau yang ditransformasikan diatur oleh aturan yang sama
- ✓ Bagaimana permintaan "penghapusan data" berlaku untuk bobot model yang dipelajari

Pertimbangkan untuk menerapkan "anggaran pelupaan"—ambang batas tentang berapa banyak data yang dapat dihafal oleh model, yang memicu pelatihan ulang ketika terlampaui. Karena tidak praktis untuk melatih ulang model AI ketika catatan subjek data tunggal telah dihapus, pendekatan ambang batas ini mungkin lebih pragmatis (dan pastikan untuk mengungkapkan praktik ini dalam NOPP).

Hak Subjek Data dalam Konteks AI

GDPR, CPRA, dan peraturan privasi lainnya telah mendefinisikan hak subjek data yang berlaku untuk semua proses manual dan otomatis, termasuk yang melibatkan AI. Mari kita lihat beberapa masalah ini.

Hak untuk Penjelasan dan Akses

Kebijakan privasi yang diperbarui harus mengklarifikasi bagaimana subjek data dapat:

- ➢ Meminta akses ke data yang digunakan untuk melatih model yang memengaruhi mereka
- ➢ Menerima penjelasan yang bermakna tentang keputusan yang didorong oleh AI

Memahami kapan sistem AI terlibat dalam pengambilan keputusan. Panduan ini selaras dengan rezim hukum seperti GDPR (Pasal 13–15, 22) dan undang-undang khusus AI yang muncul (misalnya, Undang-Undang AI Uni Eropa). Misalnya, pengguna yang ditolak permohonan hipoteknya harus diberitahu jika model AI digunakan atau memengaruhi keputusan tersebut, fitur-fitur utama yang berkontribusi terhadapnya, dan bagaimana mereka

dapat menggunakan hak mereka untuk menantang atau mengajukan banding atas keputusan tersebut.

Hak untuk Dilupakan (Penghapusan Data)

Berdasarkan bagaimana banyak model AI berfungsi dengan data pelatihan, hak untuk dilupakan mungkin menimbulkan tantangan khusus. Kebijakan privasi harus mendefinisikan apakah:

- Model harus dilatih ulang ketika data pribadi dihapus
- Penghapusan berlaku untuk sistem hilir, turunan yang disempurnakan, atau model yang dapat diakses melalui API
- Ada proses bagi subjek data untuk memverifikasi bahwa penghapusan telah terjadi

Hal ini sangat penting untuk model dasar yang dilatih pada kumpulan data web yang besar dan tidak terkurasi, mengingat risiko tinggi bahwa konten yang berbahaya, melanggar hukum, atau tidak akurat dapat dimasukkan ke dalam model AI.

Pembaruan Kebijakan Keamanan untuk Sistem AI

Organisasi perlu meninjau dan memperbarui kebijakan keamanan informasi mereka untuk mempertimbangkan penggunaan AI. Sistem bertenaga AI memperluas permukaan serangan dengan cara baru yang tidak ada dalam sistem informasi tradisional.

Memperluas Model Ancaman

Selain berbagai kerentanan tradisional, sistem AI khususnya rentan terhadap:

- Serangan adversarial yang memanipulasi input untuk menyebabkan output yang berbahaya
- Peracunan data selama pelatihan
- Pencurian model melalui akses API
- Injeksi prompt dan pembobolan sistem dalam LLM

Kebijakan keamanan tradisional seringkali mengabaikan vektor ancaman khusus AI ini. Organisasi perlu merevisi kebijakan keamanannya untuk:

- Memasukkan AI dalam latihan pemodelan ancaman
- Menentukan tanggung jawab pemantauan untuk log, telemetri, dan perilaku khusus AI
- Menerapkan pembatasan laju, kontrol akses, dan segmentasi model

Mengamankan Rantai Pasokan AI

Sistem AI bergantung pada kode sumber terbuka, model yang telah dilatih sebelumnya, dan API pihak ketiga. Kebijakan keamanan yang diperbarui harus membahas:

- Daftar Material Perangkat Lunak (SBOM) untuk pipeline AI
- Pemeriksaan dan peninjauan lisensi model pihak ketiga
- Pemindaian kerentanan berkelanjutan untuk komponen AI

Misalnya, perusahaan yang menggunakan LLM publik harus memastikan bahwa data yang diminta tidak dapat dieksfiltrasi dan bahwa LLM itu sendiri bersumber dari repositori tepercaya.

Manajemen risiko pihak ketiga dibahas lebih detail di bagian selanjutnya dalam bab ini.

Langkah-Langkah Praktis untuk Revisi Kebijakan

Mengenai revisi kebijakan keamanan dan privasi, berikut adalah pendekatan standar. Sebelum memulai bagian ini, perlu diingat bahwa saran yang disertakan di sini mengasumsikan bahwa organisasi sudah memiliki kebijakan klasifikasi data, bersama dengan inventaris data yang digunakan dalam sistem informasinya.

Organisasi yang tidak menerapkan praktik ini akan berada pada posisi yang sangat tidak menguntungkan dalam hal pengelolaan dan pemantauan data yang digunakan untuk melatih sistem AI mereka.

Lakukan Penilaian Kesenjangan

Organisasi harus memulai dengan audit kebijakan privasi dan keamanan yang berfokus pada paparan AI:

- Apakah hak subjek data diperluas ke output model?
- Apakah ada kontrol yang diterapkan untuk penggunaan kembali dan penyimpanan data pelatihan?
- Apakah ancaman AI termasuk dalam buku panduan respons insiden?

Bangun Tim Peninjau Lintas Fungsi

Peninjauan kebijakan keamanan dan privasi tidak boleh terisolasi di bidang TI, keamanan siber, atau kepatuhan. Sebaliknya, peninjauan harus mencakup:

- ✓ Petugas privasi data
- ✓ Arsitek keamanan
- ✓ Insinyur AI/ML
- ✓ Penasihat hukum dan etika

Peninjauan ini memastikan bahwa pembaruan mencerminkan realitas teknis AI dan lanskap peraturan.

Artefak Kepatuhan Kebijakan

Kebijakan keamanan yang dipublikasikan bukanlah indikasi bahwa kebijakan tersebut diikuti atau bahkan diketahui oleh personel organisasi. Sebaliknya, harus ada bukti nyata bahwa kebijakan tersebut diikuti. Kebijakan tertulis saja tidak cukup, kebijakan tersebut harus didukung oleh catatan bisnis yang membuktikan kepatuhan terhadap kebijakan. Dalam hal kebijakan teknis, konfirmasi kepatuhan mungkin melibatkan pemeriksaan arsitektur organisasi atau pengaturan konfigurasi sistem dan perangkat.

Salah satu tanda kematangan organisasi yang lebih tinggi terkait kebijakan adalah lembar kerja atau bagan yang menjelaskan lokasi atau sifat artefak yang menunjukkan kepatuhan.

3.3 KEBIJAKAN UNTUK MENGELOLA RISIKO PIHAK KETIGA

Seiring organisasi mempercepat adopsi teknologi AI, banyak komponen siklus hidup AI semakin banyak dialihdayakan, dibeli, atau didelegasikan kepada mitra eksternal. Mulai dari model yang telah dilatih sebelumnya dan infrastruktur cloud hingga annotator crowdsourcing dan API algoritmik, pihak ketiga kini memainkan peran sentral dalam pengembangan dan

pengoperasian sistem AI. Seiring dengan semakin banyaknya kemampuan AI, tata kelola harus mengikuti data dan model, ke mana pun mereka pergi.

Ekosistem yang diperluas ini menciptakan tantangan tata kelola yang unik. Kegagalan pihak ketiga tunggal, seperti pelanggaran data, model yang bias, atau praktik vendor yang tidak bertanggung jawab, dapat membahayakan integritas seluruh sistem vendor dan pelanggan yang terdampak. Mengelola risiko ini membutuhkan kebijakan risiko pihak ketiga yang kuat dan spesifik untuk AI, bukan sekadar memperluas daftar periksa vendor tradisional.

Bagian ini menjelaskan pendekatan komprehensif untuk membuat dan menerapkan kebijakan yang mengatur risiko AI pihak ketiga di seluruh pengadaan, manajemen rantai pasokan, dan interaksi tenaga kerja.

Mengapa AI Melipatgandakan Risiko Pihak Ketiga

Manajemen risiko pihak ketiga sulit dilakukan sebelum adopsi AI secara luas. Munculnya AI memperkenalkan risiko baru yang semakin memperumit praktik manajemen risiko pihak ketiga.

Di Luar Model Vendor Tradisional

Dalam perusahaan tipikal, manajemen risiko pihak ketiga (TPRM) berfokus pada paparan hukum, keuangan, dan siber: memastikan vendor tersebut mampu membayar, diasuransikan, dan patuh. Ketika diimplementasikan oleh vendor dan penyedia layanan, AI memperkenalkan bentuk risiko teknis dan etis baru, seperti:

- Algoritma kotak hitam yang tidak dapat dijelaskan
- Kumpulan data yang tidak diketahui asal atau kualitas pelabelannya
- API LLM dengan keluaran yang tidak dapat diprediksi atau filter keamanan

Risiko-risiko ini sering kali lolos dari kontrol pengadaan standar karena muncul dari perilaku model, bukan dokumentasi vendor.

Penyedia layanan kesehatan mengintegrasikan model diagnostik bertenaga AI pihak ketiga ke dalam portal pasiennya. Model tersebut berkinerja baik dalam pengujian tetapi menunjukkan bias rasial dalam penerapan karena penggunaan data pelatihan yang tidak representatif. Vendor tidak mengungkapkan masalah ini, dan penyedia tidak memiliki kebijakan yang mewajibkan validasi keadilan. Akibatnya, perusahaan harus memperluas kerangka kerja TPRM untuk memasukkan kriteria penilaian risiko khusus AI, yang berfokus pada perilaku model, silsilah data, transparansi, dan kemampuan audit.

Ketergantungan Tersembunyi dan Risiko Tingkat Bawah

Sistem AI modern sering menggunakan lapisan layanan, pustaka, dan sumber data yang saling terkait. Tim pengadaan mungkin melakukan kontrak dengan satu vendor, tanpa menyadari bahwa:

- Vendor tersebut bergantung pada tim pelabelan data subkontrak di luar negeri.
- Model tersebut dilatih pada kumpulan data yang diambil dari Internet atau pada data berhak cipta tanpa izin.
- Pustaka kode yang digunakan dalam penerapan mencakup ketergantungan yang usang atau rentan.

- ❖ Satu atau lebih pihak ketiga vendor sendiri mungkin juga menggunakan kemampuan berbasis AI yang memengaruhi organisasi.

Risiko pihak ketiga harus meluas melampaui vendor langsung ke seluruh rantai pasokan AI. Tantangan praktis bagi manajer TPRM adalah: berapa banyak tingkat ke bawah yang harus diselidiki oleh program TPRM untuk mengidentifikasi risiko dengan pihak ketiga, pihak keempat, dan seterusnya?

Tujuan Kebijakan Inti untuk Manajemen Risiko AI Pihak Ketiga

Kerangka kerja tata kelola yang efektif untuk keterlibatan AI pihak ketiga harus:

- Mengidentifikasi dan mengkategorikan titik kontak pihak ketiga dalam siklus hidup AI
- Menerapkan persyaratan uji tuntas yang mencakup risiko khusus AI
- Menentukan kewajiban kontraktual seputar penggunaan data, keadilan, keamanan, akuntabilitas, dan upaya hukum
- Menetapkan mekanisme pemantauan dan peninjauan untuk memastikan kepatuhan berkelanjutan

Tujuan-tujuan ini harus didukung oleh kebijakan yang dapat ditegakkan yang sesuai dengan ketentuan kontrak, serta alur kerja operasional dalam pengadaan, kepatuhan, keamanan, dan rekayasa.

Manajemen Kerentanan Penyedia Layanan

Hanya sedikit penyedia layanan yang mengizinkan pelanggan mereka untuk melakukan pemindaian kerentanan atau pengujian penetrasi pada layanan penyedia layanan. Sebaliknya, pelanggan harus bergantung pada audit eksternal dan validasi lain dari penyedia layanan, yang sering kali dilakukan oleh perusahaan keamanan siber yang berkualifikasi. Penyedia layanan kemudian menyediakan laporan audit dan validasi ini kepada pelanggan, biasanya dengan ketentuan kerahasiaan.

Kontrol Kebijakan Pengadaan untuk Solusi yang Didukung AI

Organisasi yang mengambil pendekatan yang lebih metodis dan matang terhadap pengadaan dapat memasukkan langkah-langkah khusus AI ke dalam prosedur penerimaan vendor dan pembelian mereka. Sangat penting untuk menentukan apakah vendor dan penyedia layanan organisasi menggunakan AI dan bagaimana mereka menggunakannya, karena beberapa aplikasi dapat mengubah profil risiko organisasi.

Penandaan Risiko AI pada Proses Penerimaan Pengadaan

Organisasi harus memodifikasi proses penerimaan pengadaan mereka untuk menandai ketika:

- Suatu produk mencakup AI, pembelajaran mesin, LLM, atau pengambilan keputusan otomatis
- Data dari organisasi akan digunakan untuk melatih model
- Vendor memproses informasi pribadi, biometrik, atau sensitif

Setiap tindakan pengadaan yang ditandai sebagai "relevan dengan AI" harus memicu tinjauan tambahan oleh pemangku kepentingan hukum, risiko, dan teknis. Contoh layanan yang relevan

dengan AI meliputi alat penyaringan resume, chatbot layanan pelanggan, dan mesin deteksi penipuan.

Uji Tuntas Khusus AI

Kebijakan pengadaan harus mewajibkan vendor untuk menyediakan:

- Dokumentasi sumber data pelatihan model dan perizinannya
- Penjelasan logika model atau kriteria pengambilan keputusan (atau justifikasi atas ketiadaannya)
- Hasil audit internal atau pihak ketiga (misalnya, untuk bias, keamanan, ketahanan)
- Bukti kepatuhan terhadap peraturan yang berlaku (misalnya, CPRA, GDPR, Undang-Undang AI Uni Eropa)

Organisasi dapat memutuskan untuk menggunakan kuesioner pemasok AI standar yang memetakan respons ke tingkat risiko dan persyaratan persetujuan.

Tingkat Risiko Vendor

Mengklasifikasikan vendor ke dalam tingkat risiko adalah praktik standar dalam TPRM, karena tidak semua vendor menimbulkan risiko yang sama. Kebijakan harus mengklasifikasikan vendor ke dalam tingkat risiko berdasarkan:

- Kritikalitas fungsi (misalnya, berdampak pada pengguna atau keputusan keuangan)
- Sensitivitas data (misalnya, PII, data kesehatan)
- Tingkat otomatisasi (misalnya, sepenuhnya otonom vs. campur tangan manusia)
- Konteks peraturan

Tingkat risiko yang lebih tinggi harus memicu:

- ✓ Tinjauan yang lebih ketat (misalnya, pengujian keadilan)
- ✓ Pengamanan kontraktual (misalnya, klausul respons insiden dan ganti rugi)
- ✓ Persyaratan pemantauan berkelanjutan

Tidak Semua Pihak Ketiga Mengungkapkan Kemampuan AI Mereka

Tidak semua pihak ketiga mengungkapkan keberadaan kemampuan AI dalam ekosistem mereka, meskipun ini bukanlah masalah hitam-putih. Sebuah vendor mungkin menggunakan AI dalam pemasaran dan jangkauannya, tetapi tidak dalam pengembangan produknya, sementara vendor lain mungkin mengizinkan pengembangnya untuk menggunakan AI untuk pembuatan kode atau pengujian kualitas. Intinya—tidak cukup hanya bertanya apakah AI digunakan dalam organisasi vendor; sangat penting untuk memahami bagaimana AI tersebut digunakan.

Pengamanan Kontrak dan Hukum

Organisasi harus memasukkan klausul dan ketentuan terkait AI ke dalam templat kontrak mereka untuk mengelola risiko di tingkat perusahaan. Beberapa vendor mungkin menggunakan terminologi terkait AI mereka sendiri, sehingga penasihat hukum perlu meninjau bahasa kontrak vendor dengan cermat.

Klausul Kontrak Khusus AI

Kontrak vendor standar seringkali tidak menyebutkan risiko khusus AI. Organisasi harus memperbarui templat kontrak standar mereka untuk memasukkan ketentuan-ketentuan berikut:

- **Persyaratan pengungkapan:** Vendor harus mengungkapkan penggunaan kemampuan AI di organisasi mereka atau oleh pihak ketiga mereka.
- **Batasan penggunaan data:** Vendor dilarang menggunakan data pelanggan untuk pelatihan ulang tanpa persetujuan eksplisit.
- **Keadilan dan bias:** Kontrak harus mewajibkan vendor untuk menguji dan memperbaiki dampak yang berbeda.
- **Praktik keamanan:** Kontrak harus mendefinisikan persyaratan untuk kontrol akses model dan pelaporan insiden.
- **Pemantauan berkelanjutan:** Vendor harus menerapkan pemantauan berkelanjutan untuk mendeteksi berbagai jenis masalah dalam sistem AI-nya.
- **Hak audit:** Kontrak harus memberikan hak kepada pelanggan untuk mengaudit atau memeriksa sistem vendor berisiko tinggi.
- **Penghentian:** Kontrak harus mendefinisikan kondisi terkait AI yang memungkinkan organisasi pelanggan untuk mengakhiri perjanjian, baik karena alasan yang sah maupun tidak.

Berikut contoh klausul tentang pelatihan model: Vendor harus mengungkapkan sumber data dan teknik pra-pemrosesan yang digunakan dalam pelatihan model dan menyatakan bahwa tidak ada data pribadi yang diperoleh tanpa dasar hukum yang sah. Beberapa organisasi akan menggunakan klausul yang diperluas dengan lebih banyak ketentuan.

Tanggung Jawab dan Penanganan Insiden

Insiden dan pelanggaran keamanan kadang-kadang terjadi di organisasi vendor dan penyedia layanan. Bahkan, organisasi tersebut dapat menjadi sasaran jika penyerang ingin menyerang pelanggan mereka. Kebijakan yang diperbarui juga harus mensyaratkan:

- Definisi tanggung jawab yang jelas jika terjadi kesalahan atau kerugian AI
- Pemberitahuan insiden yang dibatasi waktu (misalnya, dalam 24 jam)
- Hak penonaktifan model jika risiko melebihi ambang batas yang dapat diterima

Ketentuan seperti ini membantu memastikan bahwa akuntabilitas hilir dapat ditegakkan, bukan hanya sekadar aspirasi.

Tata Kelola Rantai Pasokan AI

Karena Anda membaca buku ini, Anda mungkin yakin bahwa tata kelola AI sangat penting untuk keberhasilan penggunaan kemampuan AI oleh suatu organisasi. Karena sebagian besar teknologi informasi (perangkat keras, perangkat lunak, aplikasi, dan alat) dikembangkan oleh organisasi lain, tata kelola AI perlu diperluas ke rantai pasokan dan diintegrasikan ke dalam proses manajemen risiko pihak ketiga organisasi.

Pemetaan Rantai Pasokan AI

Organisasi harus menciptakan visibilitas ke dalam rantai pasokan AI mereka untuk memberikan kesadaran berkelanjutan tentang keberadaan dan keadaan semua model dan sistem AI. Ini termasuk:

- Model yang dibangun sendiri menggunakan data atau alat pihak ketiga
- Pustaka sumber terbuka yang digunakan dalam alur kerja AI
- Model dasar yang telah dilatih sebelumnya dan disempurnakan untuk penggunaan internal

Pertimbangkan untuk mewajibkan vendor untuk mengirimkan daftar komponen sistem AI (AI-SBOM) yang mencantumkan semua input hulu (kode, data, platform, dan sebagainya) yang digunakan dalam pengembangan model.

Kebijakan Risiko AI Sumber Terbuka

Banyak tim AI bergantung pada model, kumpulan data, dan perangkat sumber terbuka. Kebijakan terkait harus membahas:

- Kejelasan lisensi: Pastikan semua aset sumber terbuka dilisensikan dengan benar.
- Pemeriksaan keamanan: Pindai paket untuk kerentanan yang diketahui.
- Tinjauan bias: Memvalidasi kualitas dan representasi data dalam kumpulan data terbuka.

Organisasi yang menggunakan sumber daya sumber terbuka dalam sistem kritis atau strategis harus mempertimbangkan untuk membentuk fungsi tata kelola sumber terbuka untuk menilai dan menyetujui alat-alat yang berdampak tinggi.

Perangkat Lunak Sumber Terbuka Ternyata Kurang Aman dari yang Kita Kira

Selama bertahun-tahun, ada pemahaman luas bahwa perangkat lunak sumber terbuka lebih aman karena kode sumbernya tersedia, memungkinkan banyak individu untuk memeriksanya guna mencari cacat. Namun, sebuah makalah tahun 2022 berjudul “Tragedi Commons Digital” karya Chinmayi Sharma membantah keyakinan yang telah lama dipegang ini.

Isu-isu AI Pihak Ketiga Lainnya

Terdapat beberapa kasus penggunaan lain di mana outsourcing dapat memengaruhi profil risiko suatu organisasi. Dua kasus tersebut dibahas di sini.

Kontrak dengan Pemberi Anotasi Manusia dan Pekerja AI

Pengembangan kemampuan terkait AI seringkali bergantung pada crowdsourcing untuk tugas-tugas seperti pelabelan data, pakar linguistik untuk penyempurnaan, dan konsultan eksternal untuk tinjauan, penilaian, dan audit. Dalam hal ini, organisasi harus mempertimbangkan kebijakan yang mewajibkan:

- Pemeriksaan praktik vendor untuk standar kerja yang etis
- Perjanjian kerahasiaan data dengan semua pemberi anotasi
- Instruksi yang jelas dan protokol QA untuk pemberi label manusia

Misalnya, sebuah perusahaan menggunakan pemberi anotasi untuk menandai konten berbahaya di situs media sosial.

Penggunaan Alat AI Generatif oleh Karyawan

Dimensi berbeda dari AI yang dialihdayakan melibatkan karyawan itu sendiri, yang dapat memperkenalkan risiko pihak ketiga dengan:

- Mengunggah data sensitif ke dalam alat seperti ChatGPT atau Gemini
- Menyalin dan menempel kode dari LLM
- Menggunakan plugin atau API yang tidak sah

Organisasi harus mengubah Kebijakan Penggunaan yang Dapat Diterima (AUP) mereka untuk mendefinisikan:

- ✓ Alat dan kasus penggunaan terkait AI yang disetujui
- ✓ Jenis data yang dilarang untuk input (kebijakan klasifikasi data mendefinisikan jenis data)
- ✓ Mekanisme pencatatan dan peninjauan untuk penggunaan AI

Organisasi biasanya melakukan program pelatihan untuk memberi tahu pekerja tentang maksud di balik kebijakan keamanan, kebijakan penggunaan yang dapat diterima, dan kebijakan relevan lainnya, serta bagaimana kebijakan ini ditegakkan.

Pemantauan dan Peninjauan Risiko AI Pihak Ketiga

Risiko pihak ketiga tidak statis. Sebaliknya, risiko yang terkait dengan penggunaan AI oleh pihak ketiga dapat berubah secara drastis, memengaruhi profil risiko organisasi. Pekerjaan proaktif diperlukan untuk mendeteksi perubahan risiko ini.

Pengawasan Vendor yang Berkelanjutan

Kebijakan manajemen risiko pihak ketiga suatu organisasi harus mewajibkan tinjauan berkala terhadap:

- ➢ Kinerja model (misalnya, penyimpangan, bias)
- ➢ Kepatuhan terhadap ketentuan kontrak
- ➢ Perubahan praktik bisnis atau kepemilikan vendor

Vendor berisiko tinggi harus ditempatkan dalam program pemantauan berkelanjutan, termasuk pemeriksaan berkala dan sertifikasi ulang tahunan.

Pemicu untuk Penilaian Ulang

Peristiwa baru harus segera memicu penilaian ulang risiko pihak ketiga yang didorong oleh kebijakan. Pemicu yang terjadi di dalam pihak ketiga mana pun dapat mencakup:

- ❖ Penambahan atau perubahan dalam layanan yang disediakan oleh pihak ketiga
- ❖ Insiden dan pelanggaran keamanan
- ❖ Kontroversi publik yang melibatkan vendor (misalnya, penyalahgunaan model)
- ❖ Insiden teknis (misalnya, halusinasi atau pemadaman)
- ❖ Perubahan peraturan yang memengaruhi domain model

Tanpa pemantauan berkelanjutan, organisasi cenderung melewatkan peristiwa yang akan memicu penilaian ulang vendor dan sistem AI mereka.

3.4 RINGKASAN

Bab ini mengkaji peran penting kebijakan dalam membangun pengawasan dan akuntabilitas di seluruh siklus hidup AI. Seiring dengan semakin kuat dan meluasnya AI, organisasi harus memastikan bahwa mekanisme tata kelola mereka tidak terbatas pada tindakan reaktif atau audit teknis yang sempit. Sebaliknya, tata kelola AI yang efektif membutuhkan kerangka kebijakan terintegrasi yang dimulai pada tahap paling awal desain sistem dan berlanjut melalui penerapan, pemantauan, dan respons insiden.

Kebijakan AI membutuhkan pendekatan siklus hidup, yang menyoroti pentingnya menciptakan titik kontak pengawasan di setiap fase pengembangan sistem utama, tidak berbeda dengan SDLC tradisional. Mulai dari penilaian kasus penggunaan dan tinjauan dampak etis hingga pelatihan, pengujian, dan pemantauan pasca-penerapan, setiap tahap menghadirkan risiko unik. Satu kesalahan langkah, seperti data pelatihan yang bias atau kasus ekstrem yang belum diuji, dapat menyebabkan kerugian yang tidak diinginkan, pelanggaran peraturan, atau reaksi publik. Kebijakan harus mendefinisikan peran, menetapkan titik pemeriksaan tinjauan, menguraikan standar dokumentasi, dan menetapkan jalur eskalasi untuk memastikan bahwa keputusan tersebut disengaja dan dapat dilacak.

Etika berdasarkan desain, tata kelola data, akuntabilitas model, dan respons insiden semuanya memainkan peran penting dalam strategi pengawasan berkelanjutan ini. Organisasi perlu mengevaluasi dan memperbarui kebijakan privasi dan keamanan data melalui lensa khusus AI. Kerangka kerja tradisional tidak dirancang untuk menangani kompleksitas yang ditimbulkan oleh model AI yang dapat menyimpulkan, menghafal, atau menggunakan kembali data dengan cara baru dan tidak terduga.

Organisasi harus menata ulang tata kelola data untuk memperhitungkan umur data yang lebih panjang dalam artefak model, risiko identifikasi ulang melalui penelusuran model, dan tantangan dalam memberikan hak subjek data (seperti penghapusan dan penjelasan) setelah data dimasukkan ke dalam model. Hal ini membutuhkan peningkatan standar anonimisasi, kontrol retensi berbasis siklus hidup, dan penerapan prinsip privasi yang lebih ketat baik dalam data terstruktur maupun output yang didorong oleh AI. Kebijakan keamanan juga harus berkembang untuk mengatasi ancaman khusus AI, seperti serangan yang merugikan, inversi model, dan rantai pasokan yang terganggu.

Organisasi perlu memperluas fokus tata kelola mereka untuk memasukkan risiko pihak ketiga, masalah yang semakin mendesak karena organisasi melakukan outsourcing pengembangan AI, pelabelan data, infrastruktur, dan alat. Praktik manajemen risiko vendor tradisional tidak memadai jika diterapkan pada sistem AI. Organisasi harus mengembangkan kebijakan pengadaan baru yang menandai sistem yang didukung AI, menuntut transparansi tentang perilaku model dan sumber data, serta menegakkan ketentuan kontrak yang mencakup keadilan, penjelasan, dan kepatuhan.

Rantai pasokan AI, termasuk model dan data sumber terbuka, harus dipetakan dan diverifikasi, sementara kebijakan tenaga kerja harus membahas baik annotator eksternal maupun karyawan internal yang menggunakan alat generatif. Pemantauan berkelanjutan, klasifikasi risiko bertingkat, dan pemicu penilaian ulang memastikan bahwa risiko pihak ketiga

tidak hanya diidentifikasi selama proses pengenalan tetapi juga dikelola secara berkelanjutan dari waktu ke waktu.

Bab ini memperkuat pesan bahwa tata kelola AI tidak dapat terisolasi, reaktif, atau statis. Sebaliknya, tata kelola AI harus dinamis, antisipatif, dan tertanam secara mendalam dalam orang, proses, dan teknologi yang mendukung sistem AI. Kebijakan tidak hanya berfungsi untuk mengurangi risiko hukum dan reputasi tetapi juga untuk membangun kepercayaan publik dan ketahanan operasional di era yang ditandai oleh sistem cerdas.

BAB 4
HUKUM PRIVASI DAN PERLINDUNGAN DATA

4.1 PEMBERITAHUAN, PILIHAN, PERSETUJUAN, DAN PEMBATASAN TUJUAN DALAM AI

Pemberitahuan, pilihan, persetujuan, dan pembatasan tujuan adalah pilar praktik dan peraturan privasi modern. Setiap sistem AI yang menggunakan data pribadi tidak hanya harus mematuhi hukum yang berlaku tetapi juga menanamkannya ke dalam desainnya. Bagian ini menerapkan prinsip-prinsip dasar dari Pasal 5, 6, 12–14, dan 21 GDPR ke dalam konteks pemrosesan yang didorong oleh AI.

Pemberitahuan

Sebagai elemen dasar perlindungan data, pemberitahuan mengacu pada pemberian informasi kepada individu tentang bagaimana data pribadi mereka akan dikumpulkan, digunakan, disimpan, dan dibagikan. Dalam AI, pemberitahuan juga harus mengungkapkan bagaimana pengambilan keputusan otomatis dapat memengaruhi individu, terutama ketika pembuatan profil atau inferensi terlibat.

Regulasi seperti GDPR (Pasal 13 dan 14) dan CPRA mewajibkan individu untuk menerima informasi yang tepat waktu dan jelas, idealnya sebelum pengumpulan data terjadi, termasuk:

- Tujuan pemrosesan
- Kategori data pribadi
- Dasar hukum pemrosesan
- Jangka waktu penyimpanan data
- Penerima data
- Apakah keputusan akan dibuat semata-mata melalui cara otomatis

Tantangan Unik dalam AI

Sistem AI memperkenalkan beberapa kompleksitas yang merusak praktik pemberitahuan tradisional:

- **Ketidaktransparan algoritma:** Sistem AI sering beroperasi sebagai "kotak hitam," sehingga sulit untuk menjelaskan bagaimana data diproses atau bagaimana keputusan diambil.
- **Inferensi:** AI dapat membuat data sensitif (misalnya, pandangan politik atau status kesehatan) dari input yang tidak sensitif, yang tidak secara eksplisit tercakup dalam sebagian besar undang-undang yang ada.
- **Perilaku dinamis:** Model AI dan ML berkembang seiring waktu, berpotensi mengubah cara data digunakan setelah pengumpulan awal.
- **Banyak sumber data:** Sistem AI sering menggabungkan data dari pihak ketiga, sensor, atau sumber web terbuka, sehingga lebih sulit untuk melacak asal usul dan memberi tahu subjek data.

- **Pengambilan keputusan otomatis:** Ketentuan yang ada (seperti Pasal 22 GDPR) sempit dan tidak sepenuhnya membahas dampak AI yang bernuansa pada bias, keadilan, dan diskriminasi.
- **Kesenjangan akuntabilitas:** Sebagian besar undang-undang mengasumsikan adanya "pengendali" dan "pengolah" yang jelas, tetapi rantai pasokan AI (penyedia data pelatihan, pengembang dan penyebar model, operator API, dan sebagainya) membuat hal ini sangat membingungkan.

Sebagai contoh, sebuah kota di AS menggunakan pengenalan wajah bertenaga AI untuk mengidentifikasi orang-orang yang dicurigai secara real-time. Warga yang berjalan melalui area publik tidak pernah secara eksplisit diberitahu tentang praktik ini. Meskipun sistem ini dimaksudkan untuk keselamatan publik, individu seringkali tidak menyadari bahwa data biometrik mereka sedang dikumpulkan dan diproses, yang merupakan kegagalan langsung dari persyaratan pemberitahuan. Perlu dicatat bahwa bentuk AI ini dilarang di Uni Eropa; hal ini dibahas dalam Bab 6.

Elemen-Elemen Kunci dari Pemberitahuan Privasi AI yang Efektif

Elemen-elemen dari pemberitahuan privasi yang efektif dan dapat dipertahankan meliputi:

- Bahasa yang lugas dan format berlapis
- Penjelasan tentang peran AI dalam pengambilan keputusan
- Kehadiran pengawasan manusia dalam pengambilan keputusan AI
- Kemampuan untuk mengakses atau menentang keputusan
- Kemampuan untuk menolak pengambilan keputusan otomatis
- Klarifikasi hak subjek data
- Pembenaran atas keadilan dan kebutuhan algoritma

Pilihan dan Persetujuan

Hukum privasi, seperti GDPR, mengharuskan subjek data diberi pilihan untuk memilih apakah suatu organisasi menggunakan informasi pribadi mereka dan memberikan persetujuan jika mereka menyetujui penggunaannya.

Mendefinisikan Persetujuan dan Variannya

Persetujuan adalah kesepakatan sukarela individu terhadap pemrosesan data mereka. Berdasarkan GDPR, persetujuan yang sah harus:

- Diberikan secara bebas
- Spesifik
- Berdasarkan informasi
- Tidak ambigu
- Dapat dicabut
- Tercatat

Dalam konteks AI, prinsip pilihan yang bermakna menjadi semakin penting, terutama di mana AI memiliki dampak signifikan pada hak-hak, seperti dalam pekerjaan, asuransi, penilaian kredit, atau peradilan pidana. Mekanisme persetujuan harus disesuaikan dengan konteks spesifik dan tingkat risiko yang terkait dengan pemrosesan data pribadi.

Persetujuan dapat mengambil beberapa bentuk:

- ✓ **Persetujuan eksplisit:** Diperlukan untuk data sensitif atau penggunaan AI berisiko tinggi (misalnya, AI terkait kesehatan)
- ✓ **Persetujuan implisit:** Seringkali tidak cukup untuk sebagian besar aplikasi AI
- ✓ **Pilihan ikut serta vs. pilihan keluar:** GDPR umumnya mensyaratkan pilihan ikut serta; beberapa yurisdiksi mengizinkan pilihan keluar dalam kondisi terbatas

Hambatan untuk Persetujuan yang Bermakna dalam AI

Memperoleh persetujuan sebagaimana dipersyaratkan oleh hukum tidak selalu mudah, terutama saat merancang proses bisnis dan sistem informasi. Hambatan meliputi:

- ➢ **Ketidakseimbangan kekuasaan:** Pengguna seringkali kurang memiliki daya tawar yang nyata, terutama ketika AI diintegrasikan ke dalam layanan penting.
- ➢ **Kurangnya pemahaman:** Sistem AI yang kompleks mengurangi pemahaman pengguna tentang apa yang mereka setujui.
- ➢ **Ketidaktransparan:** Beberapa organisasi mungkin tidak sepenuhnya mengungkapkan semua penggunaan informasi pribadi, baik karena penggunaan tersebut tidak terdokumentasi dengan baik atau karena organisasi tersebut tidak ingin mengungkapkannya.
- ➢ **Pola gelap:** Desain antarmuka pengguna yang menipu, memaksa, atau membingungkan konsumen untuk memberikan persetujuan tanpa pertimbangan yang sebenarnya.

Misalnya, aplikasi kesehatan mengumpulkan data biometrik untuk memberikan rekomendasi yang dipersonalisasi. Selama proses pendaftaran, pengguna disajikan dengan kotak centang yang menyatakan, “Saya menyetujui semua pemrosesan untuk personalisasi.” Tidak ada kontrol yang terperinci atau penjelasan yang jelas tentang penggunaan data. Hal ini gagal memenuhi standar untuk persetujuan yang terinformasi atau spesifik berdasarkan GDPR. Tantangan-tantangan ini menyoroti mengapa “pilihan yang bermakna” dalam konteks AI membutuhkan lebih dari sekadar kotak centang. Hal ini menuntut transparansi, penjelasan yang mudah diakses, dan alur persetujuan yang dirancang untuk menghormati otonomi pengguna.

Pola Gelap dan AI: Risiko Kepatuhan

Pola gelap merujuk pada desain antarmuka yang menipu yang memanipulasi pengguna untuk membuat pilihan yang tidak diinginkan atau tidak bijaksana. Dalam AI, hal ini dapat terwujud sebagai berikut:

- ❖ Persetujuan paksa melalui layanan yang dibundel
- ❖ Tombol opt-out yang tersembunyi atau tersamarkan
- ❖ Pengaturan default yang bias

- ❖ Ilusi pilihan, menyajikan banyak opsi yang semuanya mengarah pada hasil yang sama

Regulator di Uni Eropa dan Amerika Serikat semakin memperketat pengawasan terhadap praktik-praktik ini.

Pembatasan Tujuan

Pembatasan tujuan berarti bahwa data pribadi harus dikumpulkan untuk tujuan yang ditentukan, eksplisit, dan sah, dan tidak diproses lebih lanjut dengan cara yang tidak sesuai dengan tujuan tersebut. GDPR mendefinisikan pembatasan tujuan dalam Pasal 5(1)(b).

Sistem AI sering kali melampaui prinsip ini karena ketergantungannya pada kumpulan data multiguna yang besar dan pemodelan eksploratif. Model yang dilatih untuk mengoptimalkan rekomendasi produk dapat digunakan kembali kemudian untuk mendeteksi perilaku curang, penggunaan sekunder yang tidak diungkapkan kepada subjek data pada saat pengumpulan.

Penyimpangan Tujuan dan Penggunaan Kembali

Banyak organisasi kurang disiplin untuk mendokumentasikan dan melacak tujuan yang dinyatakan untuk mengumpulkan data pribadi tentang subjek data. Bahkan sebelum munculnya AI, organisasi akan mengumpulkan PII untuk satu tujuan dan kemudian menggunakan kembali PII tersebut untuk tujuan lain. Penyimpangan tujuan mengacu pada perluasan bertahap penggunaan data di luar tujuan awalnya. Hal ini sangat umum terjadi di lingkungan AI dan ML di mana data digunakan kembali di berbagai bidang:

- Fitur produk
- Penelitian dan pengembangan
- Strategi monetisasi
- Tugas prediksi yang tidak terkait

Meskipun beberapa penggunaan sekunder mungkin kompatibel dengan tujuan aslinya (misalnya, deteksi penipuan dalam layanan keuangan), banyak yang tidak, terutama ketika melibatkan data sensitif atau keputusan yang berdampak tinggi.

Tantangan Tujuan Khusus AI

Beberapa tantangan khusus AI yang terkait dengan pembatasan tujuan adalah sebagai berikut:

- **Pelatihan vs. inferensi:** Data yang dikumpulkan untuk pelatihan dapat digunakan kembali tanpa batas di seluruh pembaruan model.
- **Model pihak ketiga:** Organisasi dapat membeli atau melisensikan model yang dilatih pada kumpulan data yang tidak diketahui atau tidak diungkapkan.
- **Kesenjangan transparansi:** Pengguna jarang diberi tahu tentang jumlah tujuan hilir yang dilayani oleh suatu model.

Dalam satu kasus, sebuah perusahaan kartu kredit mengumpulkan data transaksi untuk mendeteksi dan mencegah penipuan. Kemudian, perusahaan tersebut menggunakan kumpulan data yang sama untuk membangun model yang menyimpulkan tingkat pendapatan pelanggan dan melakukan segmentasi pengguna untuk kampanye pemasaran—penggunaan

yang tidak diungkapkan dalam pemberitahuan privasi asli, yang menimbulkan kekhawatiran tentang pembatasan tujuan.

Pembatasan Tujuan dalam Sistem AI dan ML
Untuk mendukung pembatasan tujuan, organisasi harus:
○ Menyelaraskan dokumentasi dataset dengan penggunaan yang dimaksudkan ○ Memisahkan dataset untuk tugas model yang berbeda ○ Melakukan penilaian kompatibilitas untuk penggunaan sekunder ○ Memperbarui pemberitahuan privasi dengan setiap perubahan material

4.2 MINIMISASI DATA DAN PRIVASI BERDASARKAN DESAIN DALAM AI

Minimisasi data dan privasi berdasarkan desain adalah konsep dasar dalam privasi modern, dan harus diterapkan pada semua proses dan sistem, termasuk yang menggunakan kemampuan AI dan ML. Konsep-konsep ini berakar pada Pasal 5(1)(c) dan Pasal 25 GDPR, yang mengharuskan pengontrol untuk menerapkan langkah-langkah peningkatan privasi secara default dan berdasarkan desain.

Minimisasi Data

Konsep minimisasi data mengharuskan organisasi hanya mengumpulkan data pribadi yang "memadai, relevan, dan terbatas pada apa yang diperlukan" untuk tujuan yang dinyatakan (Pasal 5(1)(c) GDPR). Prinsip ini terkait erat dengan konsep kebutuhan, proporsionalitas, dan akuntabilitas.

Dalam sistem AI, minimisasi data menjadi kompleks karena:

- Kebutuhan yang dirasakan (dan nyata) akan kumpulan data pelatihan yang besar
- Sifat eksploratif dan eksperimental dari pengembangan model
- Potensi pengumpulan data berlebihan untuk mengantisipasi penggunaan di masa mendatang

Meskipun AI berkembang pesat dengan volume data, pengumpulan data yang tidak terbatas meningkatkan risiko privasi, paparan hukum dan peraturan, dan hutang teknis. Oleh karena itu, strategi minimisasi harus secara sengaja dirancang ke dalam siklus hidup AI.

Minimisasi Data di Seluruh Siklus Hidup AI

Berikut adalah beberapa cara implementasi minimisasi data:

- ✓ **Pengumpulan data:** Batasi cakupan data pribadi. Hindari pengumpulan data sensitif kecuali benar-benar diperlukan untuk tugas tersebut. Jangan mengumpulkan data berdasarkan spekulasi bahwa data tersebut mungkin dibutuhkan di kemudian hari.
- ✓ **Pra-pemrosesan dan pemilihan fitur:** Hilangkan fitur yang tidak berkontribusi pada kinerja model. Sembunyikan atau hapus bidang sensitif.
- ✓ **Pelatihan model:** Manfaatkan teknik seperti pembelajaran federasi (pendekatan pelatihan model di mana beberapa entitas melatih model sambil menjaga data tetap terdesentralisasi) atau privasi diferensial untuk melatih model tanpa memerlukan data pribadi mentah.

- ✓ **Inferensi dan penerapan:** Batasi input model ke kumpulan fitur minimal yang diperlukan untuk memberikan fungsionalitas.

Misalnya, lembaga penegak hukum menggunakan kepolisian prediktif berbasis AI untuk mengidentifikasi area berisiko tinggi. Awalnya, tim mengumpulkan nama, alamat, ras, pendapatan, dan riwayat hukuman sebelumnya. Setelah tinjauan minimisasi data, sistem dirancang ulang untuk hanya menggunakan tingkat kejahatan anonim dan stempel waktu yang dikumpulkan di tingkat lingkungan, sehingga menghilangkan pengidentifikasi langsung dan mengurangi risiko pembuatan profil.

Data Sintetis dan Minimisasi

Data sintetis adalah data yang dihasilkan secara artifisial yang mencerminkan sifat statistik dari kumpulan data nyata tanpa mengekspos informasi pribadi yang sebenarnya. Jika digunakan dengan benar, data sintetis dapat:

- ➢ Mengurangi ketergantungan pada data sensitif, termasuk PII (Informasi Identitas Pribadi)
- ➢ Mendukung pelatihan dan pengujian model
- ➢ Meningkatkan privasi dan kepatuhan

Namun, data sintetis tidak selalu menjaga privasi secara default. Jika dihasilkan dari model yang terlalu sesuai atau populasi kecil, data tersebut masih dapat menimbulkan risiko identifikasi ulang.

Menyeimbangkan Utilitas dan Minimisasi

Tim pengembangan AI sering berpendapat bahwa lebih banyak data menghasilkan model yang lebih baik. Meskipun ini bisa benar pada tahap awal, pengembalian yang semakin berkurang akan terjadi, dan nilai marginal dari data pribadi tambahan mungkin tidak membenarkan peningkatan risiko. Disiplin pembelajaran mesin yang menjaga privasi (PPML) berkembang untuk menantang pola pikir "lebih banyak lebih baik" dengan menekankan data cerdas, bukan data besar.

Minimisasi data dapat dicapai melalui dua pendekatan utama:

- ❖ Disiplin pengumpulan: Mengurangi ruang lingkup, detail, dan volume data pribadi yang dikumpulkan
- ❖ Perlindungan teknis: Mengurangi identifikasi dan paparan data yang telah dikumpulkan

Teknik minimisasi data meliputi:

- Penyamaran dan tokenisasi data
- Pseudonimisasi
- Pemrosesan di perangkat
- Analisis yang menjaga privasi (misalnya, enkripsi homomorfik, suatu bentuk enkripsi di mana komputasi dapat dilakukan pada data terenkripsi tanpa harus mendekripsinya)

Privasi berdasarkan Desain (PbD)

Privasi berdasarkan desain adalah pendekatan arsitektur dan organisasi yang menanamkan perlindungan privasi ke dalam sistem, proses, dan teknologi sejak awal, daripada menambahkannya kemudian. Awalnya dirumuskan oleh Dr. Ann Cavoukian pada tahun 1990-an, konsep ini memperoleh kekuatan hukum dalam Pasal 25 GDPR dan juga dirujuk dalam kerangka kerja seperti ISO/IEC 27701 dan NIST AI RMF.

Tujuan privasi berdasarkan desain dalam AI adalah untuk secara proaktif merekayasa sistem AI dan ML yang meminimalkan risiko bagi individu, sambil mempertahankan utilitas model dan tujuan bisnis.

Tujuh Prinsip Dasar Privasi Berdasarkan Desain

Prinsip-prinsip dasar privasi berdasarkan desain adalah:

- Proaktif, bukan reaktif; preventif, bukan perbaikan
- Privasi sebagai pengaturan default
- Privasi tertanam dalam desain
- Fungsionalitas penuh: positif-sum, bukan zero-sum
- Keamanan ujung-ke-ujung: perlindungan siklus hidup penuh
- Visibilitas dan transparansi
- Penghormatan terhadap privasi pengguna: tetap berpusat pada pengguna

Prinsip-prinsip ini harus diinterpretasikan dalam konteks unik pengembangan, penerapan, dan pemantauan AI.

Menerapkan Privasi Berdasarkan Desain dalam Pengembangan AI

Privasi berdasarkan desain dapat diterapkan pada berbagai tahap pengembangan, termasuk yang berikut ini.

Pada tahap perumusan masalah:

- Lakukan Penilaian Dampak Perlindungan Data (DPIA) untuk mengidentifikasi risiko terkait privasi sebelum desain sistem
- Nilai kebutuhan dan proporsionalitas
- Tentukan apakah kemampuan AI dapat dicapai dengan metode yang kurang mengganggu

Pada tahap pengembangan data dan model:

- ✓ Gunakan metode pelatihan yang menjaga privasi
- ✓ Terapkan alat deteksi bias dan interpretasi model
- ✓ Validasi keadilan dan penjelasan

Pada tahap penerapan dan pemeliharaan sistem:

- ➢ Batasi penyimpanan input mentah
- ➢ Aktifkan kontestasi dan umpan balik pengguna
- ➢ Terus pantau penyimpangan dan penurunan kepatuhan

Sebagai contoh, sebuah organisasi startup mengembangkan model pemrosesan bahasa alami untuk membantu peninjauan dokumen hukum. Alih-alih mengumpulkan seluruh berkas kasus dari firma hukum, tim tersebut menerapkan pemrosesan di tempat di mana dokumen tidak pernah meninggalkan lingkungan klien. Model dilatih pada data yang telah dianonimkan yang

diekstrak dengan pengawasan hukum. Pencatatan diminimalkan, dan jejak audit dienkripsi untuk meningkatkan keamanan. Langkah-langkah ini mencontohkan konsep privasi berdasarkan desain dalam praktik.

Mengintegrasikan Privasi Sejak Awal dalam Pipeline MLOps

Privasi sejak awal dapat diintegrasikan ke dalam pipeline pengembangan AI modern melalui:

- Integrasi CI/CD: Sematkan pemindaian privasi dalam proses build otomatis.
- Pemeriksaan privasi: Periksa penggunaan data atau jalur kode yang berbahaya.
- Kartu model: Dokumentasikan karakteristik privasi dan penilaian risiko.
- Deteksi penyimpangan: Pantau kapan data dunia nyata menyimpang dari data pelatihan, yang berpotensi menciptakan risiko baru.

Tim MLOps harus memperlakukan privasi sebagai metrik kinerja utama bersama dengan akurasi dan latensi.

Menyelaraskan Privasi Berdasarkan Desain dengan Peran Organisasi

Untuk mengoperasionalkan privasi berdasarkan desain dalam AI, koordinasi lintas fungsi sangat penting dan dapat mencakup kerja sama dari kelompok-kelompok berikut:

- Tim privasi harus bekerja sama dengan ilmuwan data dan insinyur MLOps.
- Manajer produk harus memperjuangkan fitur-fitur yang menghormati privasi.
- Tim hukum dan kepatuhan harus mengawasi penilaian risiko dan dokumentasi.
- Tim keamanan harus menerapkan kontrol akses, enkripsi, dan pemantauan untuk melindungi data pribadi yang sedang digunakan, dalam perjalanan, dan saat disimpan.
- Para eksekutif harus menumbuhkan budaya pengembangan AI yang bertanggung jawab.

Menanamkan privasi ke dalam pengembangan sistem AI bukan hanya masalah teknis tetapi komitmen budaya strategis yang membutuhkan investasi, pelatihan, dan dukungan kepemimpinan.

Alat dan Kerangka Kerja Praktis

Semakin banyak kerangka kerja yang mendukung integrasi privasi berdasarkan desain dalam pengembangan sistem AI:

- **ISO/IEC 2C894:** Manajemen risiko AI
- **NIST AI RMF:** Fungsi inti meliputi "memetakan, mengukur, mengelola, mengatur"
- **Pedoman ENISA:** Fokus pada permukaan serangan dan mitigasi khusus AI
- **DPIA khusus AI:** Mencakup templat yang dikembangkan oleh ICO (Inggris), CNIL (Prancis), dan lainnya

Para profesional tata kelola harus memastikan bahwa kebijakan perusahaan menetapkan bahwa proyek AI secara konsisten menggunakan alat-alat ini dan bahwa hasilnya (termasuk skor risiko, rencana mitigasi, dan risiko residual) dicatat untuk kesiapan audit.

4.3 IMPLIKASI PRAKTIS DAN TATA KELOLA

Sampai saat ini, bab ini telah membahas prinsip-prinsip privasi yang terkenal dalam konteks sistem AI. Sekarang, bab ini membahas bagaimana menerapkan prinsip-prinsip ini dalam praktik. Bagian ini mencerminkan kewajiban akuntabilitas berdasarkan Pasal 5(2) GDPR, yang diperkuat oleh Pasal 24, yang menguraikan tanggung jawab pengendali data.

AI tidak dikecualikan dari hukum privasi, dan organisasi yang paling bertanggung jawab menyadari bahwa privasi bukanlah kendala bagi inovasi, melainkan katalis untuk membangun sistem yang dapat dipercaya dan berkualitas tinggi.

Kepatuhan bukanlah aktivitas sekadar mencentang kotak. Ini adalah proses dinamis dan lintas fungsi yang harus berkembang seiring dengan sistem AI. Para profesional tata kelola harus memiliki kemampuan untuk menafsirkan mandat hukum dan ketangkasan untuk mengadaptasinya ke dalam alur kerja pembelajaran mesin, alur kerja penerapan model, dan lingkungan pengambilan keputusan secara real-time.

Seiring sistem AI membentuk kembali cara data pribadi diproses dan keputusan dibuat, penerapan prinsip-prinsip privasi klasik tetap penting dan semakin kompleks. Pemberitahuan, pilihan, persetujuan, dan pembatasan tujuan berfungsi sebagai hak-hak mendasar yang menjunjung tinggi otonomi dan kepercayaan individu. Minimalisasi data dan privasi berdasarkan desain memastikan hak-hak ini dihormati di seluruh arsitektur teknis dan praktik operasional AI.

Mengoperasionalkan Prinsip Privasi dalam AI

Menerjemahkan konsep pemberitahuan, pilihan, persetujuan, pembatasan tujuan, minimalisasi data, dan privasi berdasarkan desain ke dalam praktik yang dapat ditindaklanjuti membutuhkan struktur tata kelola yang kuat. Prinsip-prinsip ini harus dioperasionalkan di seluruh siklus hidup pengembangan dan penerapan AI, dari ide hingga penghentian model.
Hal ini membutuhkan:

- Penyematan titik pemeriksaan privasi dalam alur kerja AI
- Menyelaraskan tim hukum, teknis, dan bisnis
- Membangun mekanisme akuntabilitas dan pengawasan

Titik Sentuh Tata Kelola di Seluruh Siklus Hidup AI

Tabel 4.1 mencakup aktivitas utama yang dilakukan pada berbagai tahap pengembangan sistem AI.

Tabel 4.1	
Aktivitas Privasi dalam Siklus Pengembangan Sistem AI	
Tahap Aktivitas Privasi Siklus Hidup AI	
Definisi masalah	Melakukan DPIA (*Disclosure and Barring Service*), mendefinisikan tujuan yang sah, menilai kebutuhannya.
Akuisisi data	Validasi dasar hukum, terapkan minimalisasi data, nilai sumber pihak ketiga.
Pengembangan model	Melakukan audit bias dan keadilan, menyematkan teknik ML yang menjaga privasi, menerapkan pseudonimisasi/anonimisasi.

Evaluasi model	Konfirmasikan bahwa hasil keluaran sesuai dengan tujuan awal, validasi kemampuan penjelasan dan transparansi.
Penyebaran	Aktifkan opsi untuk menolak, berikan opsi untuk mengajukan keberatan dan pemberitahuan, terapkan kontrol akses, enkripsi data saat dalam perjalanan dan saat disimpan.
Pemantauan dan pembaruan	Pantau penyimpangan, pastikan tujuan tetap konsisten, evaluasi ulang risiko.

Para profesional tata kelola harus memastikan bahwa privasi tidak diperlakukan sebagai aktivitas sekali waktu, tetapi sebagai kewajiban berkelanjutan.

Peran dan Tanggung Jawab Privasi

Sebagai bagian dari program tata kelola AI, tanggung jawab terkait privasi harus secara formal diberikan kepada berbagai posisi, termasuk:

- **Petugas perlindungan data (DPO):** Memastikan kepatuhan hukum dan memberikan nasihat tentang DPIA, manajemen risiko, dan hak subjek.
- **Kepala petugas AI/pemimpin tata kelola AI:** Mendorong pengembangan dan penggunaan AI yang etis, memastikan keselarasan dengan kebijakan internal dan peraturan eksternal.
- **Tim MLOps dan rekayasa:** Menerapkan infrastruktur, pencatatan, pemantauan, dan manajemen siklus hidup model yang menghormati privasi untuk memastikan integritas dan keamanan data.
- **Desainer produk dan UX:** Pastikan mekanisme persetujuan dan antarmuka pengguna intuitif, adil, dan bebas dari pola gelap.
- **Tim keamanan:** Terapkan pengamanan teknis seperti enkripsi, pseudonimisasi, deteksi intrusi, dan respons insiden untuk melindungi data sistem AI.
- **Tim hukum dan kepatuhan:** Terjemahkan persyaratan hukum ke dalam kontrol praktis, lakukan audit, dan kelola respons pelanggaran.

Pertimbangan tambahan, seperti peraturan yang berlaku, sektor industri, dan budaya organisasi, mungkin memerlukan tanggung jawab tambahan.

Dokumentasi dan Kemampuan Audit

Regulator semakin mengharapkan sistem AI untuk bertanggung jawab dan mampu menunjukkan kepatuhan melalui dokumentasi dan bukti. Artefak akuntabilitas utama meliputi:

- ✓ Diagram alur data
- ✓ Kartu model dan kartu sistem
- ✓ Catatan aktivitas pemrosesan
- ✓ Log persetujuan dan keterkaitan tujuan
- ✓ Hasil penilaian risiko (DPIA, uji keseimbangan LIA)
- ✓ Laporan validasi dan pengujian

Memiliki register risiko AI terpusat atau pusat dokumentasi dapat secara signifikan meningkatkan kesiapan audit dan mendukung kewajiban transparansi.

Regulasi AI masih terus berkembang. Ketiadaan regulasi terkait AI di sektor industri atau wilayah tertentu mungkin hanya bersifat sementara. Organisasi harus berhati-hati, dengan asumsi bahwa regulasi AI mungkin akan segera diberlakukan.

Panduan Tata Kelola AI: Kebijakan Utama yang Harus Diimplementasikan

Saat mengembangkan program tata kelola AI, beberapa kebijakan yang perlu dikembangkan atau direvisi meliputi:

- Kebijakan penggunaan AI: Mendefinisikan penggunaan AI yang dapat diterima dan dilarang
- Kebijakan dampak privasi: Mewajibkan DPIA untuk sistem AI berisiko tinggi
- Kebijakan tata kelola data: Mencakup silsilah, kualitas, klasifikasi, dan minimalisasi
- Kebijakan tata kelola model: Menjelaskan pembuatan versi, pelatihan ulang, dan validasi
- Kebijakan risiko pihak ketiga: Mengatur alat AI vendor dan sumber data
- Kebijakan respons insiden untuk sistem AI: Mengintegrasikan skenario pelanggaran, kegagalan model, atau penyalahgunaan khusus AI ke dalam rencana respons insiden keamanan organisasi
- Kebijakan manajemen perubahan untuk model AI: Mendefinisikan prosedur peninjauan, persetujuan, dan pengembalian (rollback) saat memperbarui model AI.

Meningkatkan Skalabilitas Tata Kelola

Proses tata kelola dapat menjadi tantangan untuk ditingkatkan skalanya bagi organisasi yang lebih besar. Beberapa ide untuk meningkatkan skalabilitas tata kelola AI meliputi:

- Komite risiko AI atau dewan peninjau
- Tata kelola berjenjang berdasarkan tingkat risiko AI (misalnya, AI berisiko tinggi vs. otomatisasi berdampak rendah)
- Integrasi alat dengan katalog data, MLOps, dan platform tata kelola
- Program pelatihan untuk membangun kesadaran akan risiko privasi AI di seluruh organisasi
- Pusat keunggulan AI (COE) untuk mendorong kolaborasi dan pembelajaran antar tim

Tata kelola AI harus disesuaikan dengan konteks tetapi selalu berlandaskan prinsip-prinsip akuntabilitas, transparansi, dan keadilan.

4.4 KEWAJIBAN PENGONTROL DATA DALAM KONTEKS AI

Pengontrol data memainkan peran penting dalam pemrosesan informasi pribadi. Bagian ini mengeksplorasi detail kewajiban pengontrol data. Kewajiban dalam bagian ini berasal dari beberapa pasal GDPR, termasuk Pasal 5, 24–30, dan 44–49, yang mengatur tugas pengontrol, hubungan pengolah, dan transfer data lintas batas.

Sistem AI memperkenalkan risiko baru terhadap data pribadi, dan pengontrol data menanggung beban risiko ini. Pengontrol data, sebagaimana didefinisikan dalam Pasal 4(7) GDPR, adalah entitas yang “menentukan tujuan dan cara pemrosesan.” Baik menerapkan AI di

bidang perawatan kesehatan, keuangan, periklanan, atau perekrutan, organisasi yang membuat keputusan ini memikul tanggung jawab utama untuk memastikan kepatuhan terhadap undang-undang perlindungan data.

AI tidak mengubah kerangka hukum privasi inti, tetapi memperumitnya. Karena AI seringkali melibatkan pemodelan probabilistik, inferensi yang tidak transparan, dan penggunaan kembali data skala besar, kewajiban pengendali data memerlukan adaptasi. Mulai dari melakukan DPIA hingga menghormati hak subjek data dan mengelola insiden, pengendali data harus mengubah pendekatan mereka terhadap kepatuhan melalui lensa sistem algoritmik.

Bagian ini memberikan gambaran komprehensif tentang tanggung jawab hukum dan operasional yang berlaku bagi pengendali data yang menerapkan atau menugaskan sistem AI. Setiap kewajiban dieksplorasi dalam konteks risiko AI praktis dan diilustrasikan dengan contoh dan panduan tata kelola.

Penilaian Dampak Privasi dan Manajemen Risiko

Penilaian risiko adalah landasan dari setiap program privasi, yang memungkinkan penemuan proaktif risiko terkait privasi yang mungkin ditimbulkan oleh proses atau sistem baru. Ketika AI atau ML merupakan bagian dari proses bisnis, penilaian risiko menjadi lebih penting.

Peran DPIA dalam Pengembangan AI

Penilaian Dampak Perlindungan Data (DPIA) adalah proses terstruktur yang digunakan untuk mengidentifikasi, menilai, dan mengurangi risiko terhadap individu yang timbul dari pemrosesan data berisiko tinggi. Menurut Pasal 35 GDPR, pengendali data wajib melakukan DPIA ketika pemrosesan "kemungkinan akan mengakibatkan risiko tinggi terhadap hak dan kebebasan individu."

AI hampir selalu memicu ambang batas risiko tinggi tersebut, terutama ketika melibatkan:

- Pengambilan keputusan otomatis dengan dampak hukum atau dampak signifikan serupa
- Pemrosesan data kategori sensitif dalam skala besar
- Pemantauan area yang dapat diakses publik
- Pembuatan profil sistematis

Ketidakpastian, skala, dan potensi halusinasi serta bias AI meningkatkan profil risiko di luar sistem TI tradisional, sehingga DPIA tidak hanya diwajibkan secara hukum tetapi juga penting secara strategis.

Seperti aspek tata kelola AI lainnya, ketelitian dan kedalaman DPIA dapat ditentukan sebelumnya melalui tingkat risiko awal yang ditetapkan untuk sistem baru atau yang direvisi. Penilaian Dampak Perlindungan Data (DPIA) untuk sistem AI berisiko tinggi akan berisi informasi yang lebih detail daripada DPIA untuk sistem berisiko rendah.

Siklus Hidup DPIA dalam Proyek AI

DPIA harus dilakukan di awal siklus hidup pengembangan AI, idealnya pada tahap desain, dan kemudian diperbarui sepanjang proyek. Penilaian biasanya mencakup:

- Deskripsi aktivitas pemrosesan dan tujuannya
- Deskripsi data pelatihan yang akan digunakan (pada tingkat bidang)
- Deskripsi data yang akan digunakan selama operasi sistem (pada tingkat bidang)
- Daftar hukum dan peraturan terkait keamanan siber, privasi, dan AI yang berlaku
- Evaluasi kebutuhan dan proporsionalitas
- Analisis risiko dari perspektif subjek data
- Langkah-langkah untuk mengatasi atau mengurangi risiko
- Konsultasi dengan pemangku kepentingan (termasuk DPO atau regulator, jika diperlukan)

Misalnya, pengecer berencana untuk meluncurkan sistem AI segmentasi pelanggan yang menyimpulkan tingkat pendapatan dari data transaksi dan lokasi untuk menyesuaikan pemasaran. DPIA mengungkapkan risiko diskriminasi sosial ekonomi dan potensi pengucilan. Upaya mitigasi meliputi penghapusan fitur lokasi, peningkatan transparansi model, dan penyediaan opsi untuk menolak fitur tersebut.

Kapan DPIA Harus Dilakukan? (Pemicu GDPR untuk AI)
Berdasarkan GDPR, DPIA wajib dilakukan jika sistem AI melibatkan:

- Pengambilan keputusan otomatis dengan efek hukum/serupa (misalnya, lamaran pekerjaan, penilaian kredit)
- Pembuatan profil atau analitik prediktif dalam skala besar
- Data sensitif seperti biometrik, data kesehatan, atau kepercayaan agama
- Populasi rentan (misalnya, anak-anak, lansia)
- Pemantauan perilaku secara sistematis (misalnya, melalui perangkat yang dapat dikenakan atau aplikasi)

Meskipun tidak secara eksplisit diwajibkan, regulator mendorong DPIA sebagai praktik terbaik untuk semua aplikasi AI yang memproses data pribadi. DPIA awal juga dapat berfungsi sebagai tolok ukur sistem yang dapat ditinjau kemudian.

Kesalahan Umum dalam DPIA AI

Beberapa organisasi mungkin kesulitan mengadaptasi DPIA tradisional ke sistem AI karena:

- Kurangnya kejelasan tentang batasan sistem (misalnya, di mana model berakhir dan logika bisnis dimulai)
- Pembingkaian risiko yang tidak memadai (berfokus pada risiko teknis alih-alih dampak manusia)
- Kegagalan untuk meninjau kembali DPIA setelah implementasi (terutama dalam model pembelajaran berkelanjutan)

Praktik terbaik untuk DPIA AI meliputi:

- ✓ Menggunakan templat DPIA khusus AI (misalnya, dari CNIL atau ICO)
- ✓ Melibatkan tim multidisiplin (privasi, teknik, etika, UX)
- ✓ Mendokumentasikan logika model, sumber data, dan justifikasi risiko residual

Menyelaraskan dengan Kerangka Kerja yang Sedang Berkembang

Daripada merancang DPIA AI dari awal, organisasi harus memeriksa kerangka kerja lain yang mencakup penilaian gaya DPIA, termasuk:

- **NIST AI RMF:** Merekomendasikan pemetaan risiko sejak dini dan secara iteratif
- **Prinsip AI OECD:** Menekankan konsultasi pemangku kepentingan dan transparansi
- **Undang-Undang AI Uni Eropa:** Memperkenalkan penilaian kesesuaian wajib untuk sistem AI berisiko tinggi

Pengendali data harus mengantisipasi perkembangan ini dan mengintegrasikan evaluasi risiko multi-kerangka kerja jika memungkinkan.

Menggunakan Prosesor Pihak Ketiga dalam Proyek AI

Tampaknya semua hal di dunia teknologi informasi sedang dialihdayakan ke vendor dan penyedia layanan. Dalam konteks privasi dan sistem AI, banyak organisasi memilih untuk mengalihdayakan beberapa (atau banyak) aspek pemrosesan data ke penyedia layanan eksternal.

Pengontrol vs. Pemroses dalam Konteks AI

Berdasarkan Pasal 4(8) GDPR, pemroses adalah pihak mana pun yang "memproses data pribadi atas nama pengontrol." Dalam proyek AI, pemroses pihak ketiga sering dilibatkan untuk:

- Lingkungan penyimpanan atau komputasi cloud
- Layanan pelatihan model atau MLOps
- Model pra-terlatih atau model dasar dari vendor
- Layanan klasifikasi, pelabelan, dan anotasi data
- Platform AI-as-a-Service

Pengontrol harus memastikan bahwa pemroses hanya bertindak berdasarkan instruksi yang didokumentasikan dan menerapkan pengamanan teknis dan organisasi yang tepat untuk melindungi data mereka dari penyalahgunaan dan kompromi.

Kompleksitas dan ketidaktransparanan rantai pasokan AI dapat mengaburkan batas antara pengontrol dan pemroses. Misalnya, vendor AI yang menawarkan model "kotak hitam" dengan konfigurasi terbatas mungkin berfungsi lebih seperti pengontrol bersama daripada pemroses.

Persyaratan Kontraktual

Pengontrol harus memiliki kontrak yang mengikat (biasanya Perjanjian Pemrosesan Data atau DPA) dengan setiap pemroses. Menurut Pasal 28 GDPR, kontrak harus menentukan:

- Materi pokok, durasi, sifat, dan tujuan pemrosesan
- Jenis data pribadi dan subjek data
- Kewajiban pengolah data (misalnya, keamanan, kerahasiaan, subpemrosesan)
- Bantuan dengan permintaan hak dan DPIA
- Pengembalian atau penghapusan data setelah pemrosesan berakhir
- Hak audit dan inspeksi

Sebagai contoh, lembaga keuangan membuat kontrak dengan vendor pembelajaran mesin untuk menganalisis rekaman pusat panggilan untuk jaminan kualitas. Perjanjian Perlindungan

Data (DPA) mencakup klausul yang mengizinkan vendor untuk menggunakan data tersebut untuk meningkatkan modelnya sendiri, sehingga melemahkan instruksi pengontrol dan berpotensi melanggar batasan tujuan dan aturan transfer data, terutama jika vendor tersebut melatih layanan yang tidak terkait.

Daftar Periksa Uji Tuntas untuk Vendor AI

Fungsi pengadaan atau manajemen risiko pihak ketiga suatu organisasi harus menggunakan daftar periksa atau prosedur terperinci yang mencakup langkah-langkah berikut untuk uji tuntas vendor AI:

- Apakah vendor menggunakan subkontraktor untuk pengembangan model?
- Dapatkah vendor menjelaskan asal usul dan komposisi data pelatihan?
- Apakah model tersebut dapat dijelaskan atau diaudit?
- Mekanisme pencatatan dan keamanan apa yang diterapkan?
- Dapatkah vendor mendukung permintaan subjek data?
- Apakah vendor setuju untuk memproses hanya berdasarkan instruksi dan tidak menggunakan kembali data?
- Apakah vendor menjalankan kehati-hatian yang semestinya dalam keamanan siber dan privasi?
- Apakah vendor akan setuju untuk diaudit jika terjadi pelanggaran?

Uji tuntas harus mencakup pemeriksaan teknis dan hukum, bukan hanya tinjauan persyaratan layanan.

Pergeseran Prosesor dan Kesenjangan Kepatuhan

Sistem AI seringkali dinamis, artinya perilaku prosesor dapat berubah setelah penerapan. Contoh perubahan meliputi:

- Pelatihan berkelanjutan dengan data pelanggan
- Peluncuran fitur baru yang mengubah pengumpulan dan/atau pemrosesan data
- Pembaruan model yang memperkenalkan pembuatan profil

Pengontrol harus memantau prosesor sepanjang siklus hidup kontrak dan memerlukan:

- Pemberitahuan perubahan pada prosesor pihak ketiga
- Prosedur pemberitahuan perubahan
- Log audit keamanan dan fungsionalitas
- Akses ke metrik kinerja dan keadilan

Tata kelola yang kuat sangat penting dalam mencegah "penyimpangan prosesor," di mana tugas yang didelegasikan menjadi pemrosesan yang tidak terkontrol dan tidak terkelola.

Transfer Data Lintas Batas

Komponen kunci dalam hukum privasi melibatkan lokasi pemrosesan dan penyimpanan data pribadi. Transfer data lintas batas adalah hal yang umum, dan harus didokumentasikan dan dievaluasi untuk kepatuhan.

Aliran Data Global dalam Sistem AI

Ekosistem AI pada dasarnya bersifat internasional. Aktivitas lintas batas yang umum meliputi:

- ✓ Melayani pelanggan yang tinggal di luar negeri
- ✓ Melatih model di lingkungan cloud lepas pantai
- ✓ Mengakses model yang telah dilatih sebelumnya dari vendor non-lokal
- ✓ Menggunakan telemetri global atau aliran data IoT
- ✓ Mengalihdayakan pelabelan data ke wilayah dengan biaya lebih rendah

Berdasarkan GDPR, data pribadi hanya dapat ditransfer ke luar EEA jika ada perlindungan yang sesuai. Ini termasuk:

- ➢ Keputusan kecukupan (misalnya, Jepang, Korea Selatan, Inggris)
- ➢ Klausul Kontrak Standar (SCC)
- ➢ Aturan Perusahaan yang Mengikat (BCR)
- ➢ Penyimpangan untuk situasi tertentu (misalnya, persetujuan eksplisit, klaim hukum)
- ➢ Proyek AI sering mengabaikan bahwa akses dari luar EEA (bahkan jika tidak ada data yang disimpan) dihitung sebagai transfer.

Tantangan Transfer Data Khusus AI

Beberapa tantangan membuat transfer data AI sangat berisiko:

- ❖ Akses berkelanjutan vs. transfer satu kali: Model dapat mengambil data pelatihan secara berkala atau terus menerus dari sumber global.
- ❖ Risiko inferensi: Bahkan jika data mentah tetap lokal, output inferensi dapat mengungkapkan informasi sensitif.
- ❖ Hukum lokalisasi data: Negara-negara seperti Tiongkok dan India memberlakukan hukum lokalisasi data yang relevan dengan AI yang mungkin bertentangan dengan aturan transfer GDPR.

Dalam satu contoh, sebuah perusahaan asisten suara yang berbasis di Uni Eropa menggunakan layanan transkripsi yang berbasis di Amerika Serikat untuk memproses perintah audio. Klausul Kontrak Standar (SCC) telah berlaku, tetapi pengungkapan data pribadi warga negara Uni Eropa di Amerika Serikat kemudian menimbulkan masalah berdasarkan Schrems II dan peraturan transfer data Uni Eropa.

Inferensi AI sebagai Transfer?

Interpretasi hukum terbaru menunjukkan bahwa:

- Akses jarak jauh ke data = transfer
- Inferensi model waktu nyata dari server asing = potensi transfer
- Penggunaan API AI pihak ketiga (misalnya, pengenalan gambar) = transfer jika data pribadi terlibat

Pengontrol harus dengan cermat menilai arsitektur layanan AI untuk menentukan apakah ada risiko lintas batas—bahkan ketika data "tetap di tempatnya."

Penilaian Dampak Transfer (*Transfer Impact Assessment*/TIA)

Sejak putusan Schrems II membatalkan Privacy Shield Uni Eropa-AS, organisasi harus melakukan Penilaian Dampak Transfer (TIA) ketika mengandalkan SCC atau BCR. TIA harus mempertimbangkan:

- Lingkungan hukum negara tujuan
- Hukum privasi yang berlaku
- Sifat data dan pemrosesannya
- Keberadaan pengawasan atau akses pemerintah
- Langkah-langkah tambahan (misalnya, enkripsi, pseudonimisasi)

Untuk sistem AI, TIA mungkin juga perlu membahas:

- Pelatihan model jarak jauh menggunakan data EEA
- Akses vendor ke telemetri atau log kerusakan
- Penggunaan model pra-terlatih yang dibangun di atas data non-EEA
- Panggilan API lintas batas atau permintaan inferensi yang mengirimkan data pribadi ke server non-EEA
- Layanan hosting atau orkestrasi pihak ketiga yang menjalankan model AI pada infrastruktur di luar EEA

Pengontrol harus mendokumentasikan TIA dan siap untuk membenarkan keputusan mereka kepada otoritas pengawas.

Hak Subjek Data dan AI

Berdasarkan GDPR dan kerangka kerja serupa (misalnya, CPRA, LGPD), subjek data memiliki serangkaian hak, termasuk:

- Hak akses (Pasal 15 GDPR)
- Hak untuk perbaikan (Pasal 16 GDPR, CPRA 1798.106, Pasal 18(III) LGPD)
- Hak untuk penghapusan ("hak untuk dilupakan," Pasal 17 GDPR, CPRA 1798.105, Pasal 18(IV) LGPD)
- Hak untuk menolak pengambilan keputusan otomatis (CPRA 1798.121) Hak untuk pembatasan pemrosesan (Pasal 18 GDPR)
- Hak untuk portabilitas data (Pasal 20 GDPR, Pasal 18(V) LGPD) Hak untuk keberatan (Pasal 21 GDPR)
- Hak untuk tidak tunduk pada keputusan otomatis semata dengan dampak signifikan (Pasal 22 GDPR)

Pengontrol harus memastikan hak-hak ini tetap dapat ditindaklanjuti, bahkan ketika keputusan dibantu oleh sistem AI yang kompleks.

Tantangan Implementasi Khusus AI

Dalam sistem AI, hak-hak ini menimbulkan tantangan teknis dan operasional:

- ✓ **Akses dan penjelasan:** Sulit untuk menjelaskan keluaran dari model kotak hitam, apalagi memberikan penjelasan yang mudah dipahami kepada subjek data.
- ✓ **Koreksi dan penghapusan:** Bagaimana Anda mengoreksi atau menghapus data yang tertanam dalam bobot model?

- ✓ **Portabilitas:** Keluaran AI seringkali bergantung pada model dan tidak mudah dipindahkan dengan cara yang berarti.
- ✓ **Keberatan terhadap pembuatan profil:** Pembuatan profil merupakan hal yang melekat pada banyak sistem AI, yang membutuhkan mekanisme penolakan yang kuat.
- ✓ **Pembatasan pemrosesan:** Menghentikan penggunaan data pribadi dalam model aktif dapat menjadi tantangan operasional, terutama untuk sistem pembelajaran berkelanjutan.

Misalnya, aplikasi berbagi tumpangan menggunakan model AI untuk menonaktifkan pengemudi berdasarkan prediksi risiko penipuan mereka. Pengemudi menerima pemberitahuan otomatis dan meminta detail. Perusahaan tersebut mengutip "algoritma kepemilikan" tanpa memberikan penjelasan yang berarti, sehingga melanggar Pasal 15 dan 22 GDPR.

Pasal 22 GDPR: Keterlibatan Manusia atau Sekadar Hiasan?

Pasal 22 melarang keputusan "yang semata-mata didasarkan pada pemrosesan otomatis" yang secara signifikan memengaruhi individu, kecuali:

- ➢ Hal itu diperlukan untuk suatu kontrak
- ➢ Hal itu diizinkan oleh hukum
- ➢ Hal itu didasarkan pada persetujuan eksplisit

Bahkan dengan peninjauan manusia, regulator memerlukan pengawasan yang nyata—bukan sekadar persetujuan tanpa pertimbangan. Pengontrol data harus memastikan bahwa peninjau manusia dapat mengesampingkan keputusan AI dan memahami dasar intervensi ini.

Solusi Praktis

Organisasi yang menghadapi tantangan hak subjek data memiliki beberapa solusi yang tersedia. Beberapa solusi ini terkait erat dengan desain sistem AI dan tidak dapat dengan mudah diterapkan kemudian.

- ❖ **Desain untuk hak:** Rancang dan bangun kemampuan menjelaskan dan pencatatan ke dalam model sejak awal.
- ❖ **Dokumentasi model:** Pertahankan metadata tentang input, versi, logika, dan dampaknya.
- ❖ **Penjelasan berlapis:** Gabungkan ringkasan tingkat tinggi dengan detail teknis jika diperlukan.
- ❖ **Anonimisasi dan pseudonimisasi:** Latih model AI dengan data yang dianonimkan atau dipseudonimkan jika memungkinkan.
- ❖ **Strategi penghapusan:** Manfaatkan alur pelatihan ulang yang mengecualikan catatan yang dihapus atau membatasi data pribadi dalam kumpulan data pelatihan.

Menghormati hak hukum dan hak asasi manusia dalam AI dapat dicapai, tetapi harus disengaja dan dimasukkan ke dalam desain awal sistem AI. AI tidak memiliki intuisi moral; ia hanya mengikuti aturan dan tujuan yang ditentukan oleh manusia. Ia tidak secara inheren

mengetahui apa yang "benar," jadi dengan mengaitkan konsep "hak" dengan hak-hak yang didefinisikan secara hukum dan diakui secara universal (misalnya, kerangka kerja hak asasi manusia dan hukum privasi) menjadikannya objektif, dapat ditegakkan, dan konsisten di berbagai konteks.

Manajemen Insiden dan Pemberitahuan Pelanggaran

Kemampuan manajemen insiden yang formal dan terdokumentasi merupakan elemen inti dari setiap program keamanan siber dan privasi. Bagian ini membahas tambahan dan pertimbangan terkait AI untuk manajemen insiden dan pelanggaran. Untuk tinjauan komprehensif tentang manajemen insiden keamanan siber dan privasi, lihat panduan studi saya tentang sertifikasi CISM (*Certified Information Security Manager*), CRISC (*Certified in Risk and Information Systems Control*), CIPM (*Certified Information Privacy Manager*), dan CDPSE (*Certified Data Privacy Solutions Engineer*).

Meninjau Kembali Pelanggaran dalam Sistem AI

Sistem AI memperkenalkan jenis pelanggaran keamanan dan privasi baru, termasuk:

- **Kebocoran data pelatihan:** Misalnya, melalui serangan inversi atau inferensi keanggotaan, atau serangan ekstraksi gradien
- **Kebocoran output:** Misalnya, LLM mereproduksi data pelatihan secara verbatim atau mengungkapkan konten sensitif melalui injeksi prompt
- **Amplifikasi bias:** Menghasilkan keputusan yang merugikan dan tidak adil dalam skala besar
- **Pergeseran atau peracunan model:** Merusak kualitas keputusan melalui input yang merugikan

Insiden keamanan tradisional (misalnya, pencurian kredensial, malware) masih berlaku, tetapi pengontrol harus memperluas deteksi pelanggaran untuk mencakup vektor khusus AI.

Persyaratan Pemberitahuan Pelanggaran GDPR

Menurut Pasal 33 GDPR, pengontrol harus memberi tahu otoritas pengawas dalam waktu 72 jam setelah mengetahui adanya pelanggaran data pribadi. Selain itu, jika ada risiko tinggi terhadap hak dan kebebasan subjek data, mereka harus diberitahu tanpa penundaan yang tidak semestinya. Pemberitahuan tersebut harus mencakup:

- Sifat dan cakupan pelanggaran
- Kategori dan jumlah subjek data yang terpengaruh
- Konsekuensi yang mungkin terjadi
- Langkah-langkah yang diambil atau diusulkan untuk mengurangi kerugian

Kegagalan untuk mematuhi dapat mengakibatkan denda hingga 310 juta atau 2% dari omset global perusahaan, dan berpotensi hukuman yang lebih tinggi jika digabungkan dengan pelanggaran lainnya.

Misalnya, sebuah laboratorium penelitian menerbitkan demo AI yang dilatih pada forum kesehatan online yang di-scrape. Selama pengoperasian, model tersebut menghasilkan postingan pengguna yang sensitif hampir sama persis. Setelah diidentifikasi, pengontrol harus menilai apakah ini merupakan pelanggaran dan apakah pengguna yang terpengaruh dapat diidentifikasi dan diberitahu.

Panduan Penanganan Insiden AI yang Penting

Setiap pengendali yang menerapkan AI harus memelihara rencana dan panduan respons insiden yang sadar AI. Termasuk:

- Inventaris model AI, asal usul data, dan silsilah data
- Kriteria penilaian dampak pelanggaran untuk keluaran inferensi
- Titik kontak untuk vendor atau subprosesor
- Jalur eskalasi hukum untuk DPIA dan tinjauan Pasal 33
- Pelatihan untuk tim respons tentang vektor khusus AI

Semakin cepat tim Anda dapat membedakan bug model dari pelanggaran data pribadi, semakin baik postur kepatuhan Anda.

Deteksi dan Kesiapan Insiden AI

AI mempersulit deteksi insiden dalam beberapa cara:

- **Kegagalan diam-diam:** Output model dapat menurun tanpa tanda-tanda yang jelas.
- **Kurangnya kemampuan observasi:** Pencatatan yang buruk atau alur kerja yang tidak transparan menghambat investigasi.
- **Data tidak terstruktur:** Pelanggaran dapat memengaruhi dokumen, media, atau embedding yang tidak dipantau oleh alat DLP standar.
- **Ketergantungan pihak ketiga:** Insiden keamanan pada penyedia model hulu, API, atau platform hosting dapat membocorkan data pribadi tanpa memicu peringatan internal.

Untuk mengatasi risiko ini, pengontrol harus:

- ✓ Menerapkan pencatatan dan deteksi anomali pada lapisan model dan data.
- ✓ Menggunakan kartu model dan kartu sistem yang mendokumentasikan sumber data dan risiko.
- ✓ Menetapkan latihan pelanggaran data yang mencakup skenario AI.
- ✓ Melakukan penilaian risiko berkala, termasuk penilaian risiko pihak ketiga, untuk mengidentifikasi risiko tambahan.

Pencatatan dan Akuntabilitas

Kekuatan dan integritas program tata kelola apa pun bergantung pada dokumentasi yang efektif, pencatatan yang menyeluruh, dan akuntabilitas yang kuat.

Pengendali Harus Mendemonstrasikan Kepatuhan

Prinsip akuntabilitas GDPR (Pasal 5(2)) mengharuskan pengendali tidak hanya mematuhi prinsip-prinsip perlindungan data tetapi juga mendemonstrasikan kepatuhan tersebut. Demonstrasi ini mencakup pemeliharaan catatan internal terperinci (untuk setiap sistem AI yang digunakan) tentang:

- ➢ Tujuan pemrosesan
- ➢ Kategori subjek data dan data pribadi
- ➢ Penerima dan transfer
- ➢ Periode penyimpanan

- Langkah-langkah keamanan dan pengamanan

Untuk sistem AI, kewajiban pencatatan ini menjadi jauh lebih kompleks, terutama ketika model bersifat dinamis, diperbarui secara berkala, atau tertanam dalam produk dan layanan lain.

Organisasi yang menggunakan tingkatan risiko untuk sistem AI dapat mensyaratkan pencatatan yang lebih ketat pada sistem AI berisiko tinggi, dengan persyaratan yang kurang ketat pada sistem berisiko rendah.

Apa yang Harus Didokumentasikan dalam Proyek AI

Untuk memenuhi kewajiban akuntabilitas, pengontrol harus memelihara artefak yang menjelaskan:

- **Definisi tujuan:** Mengapa AI dikembangkan dan keputusan apa yang didukungnya
- **Asal usul data pelatihan:** Dari mana data berasal, bagaimana data dibersihkan, dan dasar hukumnya
- **Arsitektur dan logika model:** Jenis model apa yang digunakan dan bagaimana fungsinya
- **Hasil evaluasi:** Skor akurasi, keadilan, bias, ketahanan, dan penjelasan
- **Penilaian risiko:** DPIA, TIA, dan justifikasi untuk setiap risiko residual
- **Riwayat versi dan log perubahan:** Untuk pembaruan model atau perubahan kebijakan
- **Tinjauan dan persetujuan:** Siapa yang berpartisipasi dalam tinjauan gerbang pengembangan sistem, dan siapa yang menyetujui proyek pengembangan sistem untuk dilanjutkan

Misalnya, sebuah bank membangun model penilaian kredit yang diperbarui setiap kuartal. Bank tersebut memelihara registri AI yang berisi cuplikan data pelatihan setiap versi, asumsi algoritmik, teknik mitigasi bias, dan pengujian output. Ketika diaudit oleh regulator, dokumentasi tersebut mendukung klaim keadilan dan proporsionalitas.

Apa yang Harus Ada dalam Buku Catatan Kepatuhan AI Anda

Organisasi yang membangun persyaratan pencatatan tata kelola AI mereka dapat menyertakan persyaratan berikut:

- Kartu model (ringkasan model AI, risiko, penggunaan yang dimaksudkan)
- Kartu sistem (dokumentasi sistem AI ujung-ke-ujung)
- Catatan aktivitas pemrosesan (Pasal 30 GDPR)
- Catatan peristiwa, insiden, dan pelanggaran
- Log persetujuan, penolakan, dan penanganan keberatan
- DPIA, TIA, dan catatan konsultasi
- Dokumentasi proses peninjauan manusia untuk keputusan Pasal 22 GDPR

Aset-aset ini penting tidak hanya untuk regulator tetapi juga untuk audit internal, dewan etik, dan keputusan produk di masa mendatang.

Alat dan Kerangka Kerja untuk Tata Kelola

Banyak organisasi menyederhanakan pencatatan program tata kelola mereka dengan alat dan kerangka kerja, termasuk:

- **Platform tata kelola:** Seperti OneTrust, Navex, Collibra, atau BigID
- **Integrasi MLOps:** Sertakan pencatatan privasi dalam pipeline CI/CD
- **Alat dokumentasi terbuka:** Seperti Model Cards++ atau lembar data untuk kumpulan data
- **Daftar risiko AI:** Lacak risiko yang diketahui, mitigasi, dan siklus peninjauan

Menyematkan dokumentasi dalam proses pengembangan, daripada memperlakukannya sebagai pertimbangan kepatuhan setelahnya, mendukung tata kelola berkelanjutan.

4.5 PERSYARATAN YANG BERLAKU UNTUK KATEGORI DATA SENSITIF ATAU KHUSUS

Selain PII standar, mungkin ada persyaratan tambahan untuk kategori data pribadi tertentu, termasuk perawatan kesehatan, biometrik, orientasi seksual, dan afiliasi agama. Bagian ini membahas beberapa kasus khusus ini. Pemrosesan data sensitif dalam sistem AI diatur oleh Pasal 9 GDPR, bersama dengan ketentuan terkait dalam CPRA, LGPD, dan kerangka kerja global lainnya.

Sistem AI membutuhkan banyak data, dan beberapa data tersebut lebih penting daripada yang lain. Ketika sistem AI memproses kategori khusus data pribadi, taruhannya lebih tinggi, risikonya lebih tajam, dan ekspektasi regulasinya lebih ketat.

Di bawah hukum seperti GDPR Uni Eropa, CCPA/CPRA California, LGPD Brasil, dan lainnya, jenis data pribadi tertentu dianggap sensitif secara inheren. Ini termasuk data yang mengungkapkan asal ras atau etnis, opini politik, keyakinan agama atau filosofis, keanggotaan serikat pekerja, data genetik, data biometrik untuk tujuan identifikasi, data kesehatan, kehidupan seks, orientasi seksual, dan banyak lagi.

Dalam konteks AI, kategori data ini sering diproses dalam skala besar dan terkadang disimpulkan daripada dikumpulkan secara eksplisit, sehingga meningkatkan beban kepatuhan dan kewajiban etis pada pengembang, penyebar, dan pengontrol data.

Bagian ini membahas persyaratan hukum, operasional, dan tata kelola yang terkait dengan data sensitif dalam sistem AI. Ini mencakup dasar hukum untuk pemrosesan, kategori spesifik seperti biometrik dan data kesehatan, risiko terkait, perlindungan teknis, dan kewajiban berbasis hak.

Landasan Hukum dan Regulasi

Bagian ini membahas bagaimana GDPR dan peraturan lainnya memperlakukan kelas khusus data sensitif.

Pasal 9 GDPR dan Batasan Larangan

Pasal 9 GDPR menetapkan nada dengan jelas ketika menyatakan, “Pengolahan data pribadi yang mengungkapkan asal ras atau etnis, opini politik, keyakinan agama atau filosofis… dilarang.”

Perhatikan bahwa ini adalah posisi standar: pengolahan data kategori khusus dilarang kecuali salah satu dari daftar pengecualian yang terbatas berlaku. Pengecualian ini meliputi:

- Persetujuan eksplisit
- Hukum ketenagakerjaan dan perlindungan sosial
- Kepentingan vital (misalnya, keadaan darurat medis)
- Kepentingan publik di bidang kesehatan masyarakat
- Pengarsipan, penelitian, atau tujuan statistik (dengan perlindungan)
- Pembentukan atau pembelaan klaim hukum

Ketika sistem AI menyentuh kategori ini, bahkan secara tidak langsung, pengontrol harus memastikan dasar hukum yang tepat, ditambah perlindungan teknis dan organisasi untuk melindungi data.

Persamaan Internasional Utama

Selain GDPR, yurisdiksi lain juga memberlakukan kewajiban yang lebih tinggi terhadap data sensitif. Beberapa contoh dikutip dalam Tabel 4.2.

Tabel 4.2
Peraturan Privasi Terpilih Mengenai Data Sensitif

Wilayah	Hukum	Definisi Data Sensitif	Ketentuan Utama
California	CPRA	Biometrik, kesehatan, ras, agama, persatuan, geolokasi tepat	Membutuhkan pemberitahuan, opsi untuk menolak, penggunaan terbatas.
Brazil	LGPD	Biometrik, kesehatan, asal ras atau etnis, kepercayaan agama/filosofis, opini politik, keanggotaan serikat pekerja	Landasan hukum terpisah, perlindungan yang lebih ketat
Kanada	PIPEDA (diterjemahkan melalui OPC)	Kesehatan, biometrik, etnisitas, kepercayaan agama	Persetujuan harus bermakna; biometrik berisiko tinggi.
India	UU DPDP (2023)	Kesehatan, biometrik, keuangan, genetik, kehidupan seks, keyakinan politik	Berbasis persetujuan, dengan aturan khusus AI yang diantisipasi di masa mendatang.

Apa yang Termasuk dalam Kategori Data "Sensitif"?

Meskipun definisinya bervariasi, berikut ini umumnya dianggap sensitif di berbagai yurisdiksi:

- Biometrik (untuk identifikasi)
- Data kesehatan dan genetik
- Asal ras atau etnis
- Keyakinan agama atau filosofis
- Opini dan afiliasi politik
- Keanggotaan serikat pekerja
- Kehidupan seks atau orientasi seksual
- Geolokasi yang tepat (di beberapa yurisdiksi)

Beberapa undang-undang juga memperlakukan data keuangan, identitas yang dikeluarkan pemerintah, atau data anak-anak sebagai data sensitif, tergantung pada konteksnya.

Profil Risiko Unik AI untuk Data Sensitif

AI dapat memproses data sensitif secara sengaja, tidak sengaja, atau secara inferensial:

- ✓ **Sengaja:** Alat AI kesehatan menggunakan rekam medis elektronik.
- ✓ **Tidak sengaja:** Alat penyortiran CV mengumpulkan afiliasi politik pelamar dari aktivisme masa lalu.
- ✓ **Inferensial:** Model analisis wajah memprediksi etnis atau emosi tanpa diberitahu secara eksplisit.

Masing-masing skenario ini membawa risiko:

- ➢ Kurangnya dasar hukum yang sah untuk pengumpulan dan pemrosesan
- ➢ Diskriminasi dan bias
- ➢ Kurangnya transparansi
- ➢ Pengawasan dan efek menghambat (ketika orang-orang dihalangi untuk berpartisipasi dalam kegiatan legal)

Pengendali data harus memahami risiko ini dan mengambil langkah-langkah proaktif, terutama ketika menerapkan sistem berisiko tinggi yang dapat memengaruhi lapangan kerja, perawatan kesehatan, penegakan hukum, atau akses ke layanan penting.

Penilaian dampak pemrosesan data (DPIA) yang efektif mengantisipasi jenis risiko ini.

Biometrik dan Kategori Data Sensitif Berisiko Tinggi Lainnya

Kategori data khusus, sebagaimana didefinisikan dalam Pasal 9 GDPR, yang disebut sebagai "data sensitif," menerima perlindungan yang lebih tinggi karena potensinya untuk menyebabkan kerugian atau diskriminasi yang signifikan terhadap subjek data jika disalahgunakan. Jenis data ini dapat mengekspos individu pada pengawasan, pengucilan, pelecehan, atau manipulasi, terutama ketika diproses oleh sistem AI dalam skala besar.

Sistem AI tidak hanya mengonsumsi data sensitif; mereka juga dapat menghasilkannya; misalnya, dengan menyimpulkan ras, kondisi kesehatan, atau pandangan politik dari input yang tidak berbahaya, seperti perilaku penelusuran, nada suara, atau riwayat pembelian.

Bagian ini mengeksplorasi kategori data sensitif utama yang paling sering terlibat dalam pengembangan dan penerapan AI: biometrik, data kesehatan dan genetik, agama dan kepercayaan, opini politik, dan atribut demografis atau terkait tenaga kerja seperti ras dan keanggotaan serikat pekerja.

Biometrik

Data biometrik mengacu pada data pribadi yang dihasilkan dari pemrosesan teknis spesifik yang terkait dengan karakteristik fisik, fisiologis, atau perilaku seseorang yang dapat diamati atau diukur yang memungkinkan atau mengkonfirmasi identifikasi unik.

Jenis data biometrik umum meliputi:

- ❖ Fitur wajah
- ❖ Sidik jari dan sidik telapak tangan
- ❖ Pemindaian iris dan retina
- ❖ Sidik suara
- ❖ Gaya berjalan dan postur tubuh
- ❖ Ritme mengetik dan pergerakan mouse

Sistem biometrik umumnya digunakan untuk:

- Otentikasi (misalnya, masuk ke sistem atau membuka pintu)
- Identifikasi (misalnya, mencocokkan wajah dengan daftar pantauan)
- Pengawasan dan analitik (misalnya, memperkirakan usia atau emosi di ruang publik)

Data biometrik sangat sensitif karena:

- Tidak mudah diubah jika disalahgunakan (Anda bisa mendapatkan nomor kartu kredit baru, tetapi Anda tidak bisa mendapatkan sidik jari baru)
- Sering dikumpulkan secara pasif atau tanpa persetujuan
- Dapat digunakan kembali untuk pengawasan, pembuatan profil, atau manipulasi perilaku

Misalnya, pusat perbelanjaan menggunakan pengenalan wajah untuk mengidentifikasi pencuri berulang. Sistem yang sama diam-diam digunakan untuk melacak pola pergerakan pelanggan dan memicu iklan yang ditargetkan. Pelanggan tidak diberi tahu atau diberi kesempatan untuk menolak. Contoh ini melibatkan perlindungan data biometrik, pembatasan tujuan, dan kewajiban transparansi.

Apakah Semua Sistem Biometrik Berisiko Tinggi?

Tidak selalu. GDPR hanya menerapkan perlindungan yang lebih tinggi pada data biometrik yang digunakan untuk mengidentifikasi seseorang secara unik. Sistem biometrik yang digunakan semata-mata untuk kategorisasi (misalnya, estimasi kelompok usia tanpa dikaitkan dengan identitas) mungkin tidak memicu Pasal 9. Namun, hal itu masih dapat menimbulkan risiko etika dan hukum, tergantung pada konteksnya.

Data Kesehatan dan Genetik

AI mentransformasi layanan kesehatan, mulai dari diagnostik prediktif dan pengobatan personal hingga optimasi operasional dan triase pasien. Namun, data kesehatan dan genetik termasuk di antara bentuk informasi pribadi yang paling sensitif.

Kekhawatiran utama meliputi:

- Diskriminasi medis dan pekerjaan
- Penentuan kondisi tanpa persetujuan
- Pengidentifikasian ulang dari kumpulan data anonim
- Penggunaan data pasien untuk penelitian atau pelatihan yang tidak terkait

Data genetik memperkenalkan kompleksitas tambahan karena implikasi keluarga dan keturunannya, yang berpotensi memengaruhi individu yang tidak pernah memberikan persetujuan mereka.

Contoh platform telehealth menggunakan pemrosesan bahasa alami untuk menyaring pasien yang mengalami depresi. Platform ini menyimpan transkrip dan menggunakannya untuk menyempurnakan model, termasuk layanan analisis sentimen yang ditawarkan kepada perusahaan asuransi. Pasien tidak diberitahu bahwa interaksi mereka akan digunakan dengan cara ini, sehingga menimbulkan kekhawatiran tentang penggunaan data kesehatan dan potensi aplikasi sekundernya, seperti pemeriksaan pekerjaan dan kredit.

Kompromi Privasi dalam Terapi Digital

Terapi digital dan alat AI yang berfokus pada kesehatan menawarkan manfaat klinis tetapi juga menciptakan vektor paparan baru:

- Aplikasi dapat membocorkan lokasi atau metadata perangkat
- Penyimpanan cloud mungkin kurang memiliki kontrol keamanan dan privasi tingkat perawatan kesehatan
- Berbagi data dengan pihak ketiga (misalnya, pengiklan atau pelatih model) mungkin tidak dipantau

Organisasi harus selalu memperlakukan data "kesehatan konsumen" dengan kehati-hatian yang sama seperti data medis yang diatur.

Kepercayaan Agama dan Filosofis

Sistem kepercayaan, baik agama, filosofis, atau ideologis, secara eksplisit dilindungi berdasarkan GDPR. Sistem AI dapat menemukan data ini melalui:

- ✓ Tanggapan survei atau profil
- ✓ Media sosial
- ✓ Analisis tekstual (misalnya, khotbah, doa, atau kosakata keagamaan)
- ✓ Inferensi berdasarkan perilaku (misalnya, perjalanan liburan, pembelian makanan)

Memproses informasi ini tanpa dasar hukum yang kuat dapat melanggar Pasal 9 GDPR dan berisiko:

- ➢ Pembuatan profil atau stereotip
- ➢ Bias dan diskriminasi
- ➢ Efek yang menghambat kebebasan berekspresi
- ➢ Iklan yang ditargetkan atau manipulasi

Contoh model bahasa yang disempurnakan pada unggahan publik mulai menugaskan pengguna ke "kelompok kepercayaan yang mungkin." Layanan streaming menggunakan ini untuk merekomendasikan konten bertema keagamaan, dan kemudian untuk membatasi atau mengubah rekomendasi tersebut. Pengguna seringkali tidak menyadari bahwa identitas spiritual mereka telah disimpulkan dan ditindaklanjuti.

Kepercayaan pada Kumpulan Data: Risiko Tersembunyi

Meskipun kepercayaan agama tidak dikumpulkan secara eksplisit, kepercayaan tersebut dapat tertanam dalam data pelatihan. Pengembang harus mempertimbangkan:

- ❖ Bias dalam sistem moderasi konten
- ❖ Output AI yang mencerminkan atau menekan ekspresi berbasis kepercayaan
- ❖ Pelabelan yang salah atau generalisasi berlebihan terhadap atribut yang berkaitan dengan keyakinan

Audit bias harus mencakup pemeriksaan terhadap stereotip, diskriminasi, dan marginalisasi agama.

Opini dan Afiliasi Politik

Data politik sering kali disimpulkan dari perilaku pengguna, terutama dalam periklanan, media sosial, dan personalisasi konten. Kategori ini meliputi:

- Keanggotaan partai
- Riwayat pemungutan suara (spesifik atau tersirat)
- Pandangan politik yang dinyatakan
- Keterlibatan dengan konten politik

Penyalahgunaan data politik dalam sistem AI dapat:

- Memanipulasi opini (misalnya, penargetan mikro terhadap pemilih)
- Menekan perbedaan pendapat
- Mempengaruhi hasil pemilu
- Menciptakan diskriminasi berbasis profil

Contoh sistem AI digunakan oleh platform media sosial untuk menilai kecenderungan politik pengguna untuk iklan yang ditargetkan. Sistem tersebut belajar untuk menghubungkan klik artikel berita dengan ideologi tertentu dan mulai mengkurasi konten sesuai dengan itu. Gelembung filter politik meningkat, dan beberapa pengguna menjadi sasaran informasi yang salah.

Data Politik dan Otomatisasi Pengambilan Keputusan

Pengambilan keputusan otomatis berdasarkan pandangan politik, seperti kelayakan iklan, layanan keuangan, atau penyaringan karyawan, kemungkinan dilarang berdasarkan Pasal 22 GDPR kecuali jika persetujuan eksplisit diberikan dan persyaratan Pasal 9 untuk data kategori khusus juga dipenuhi. Dalam praktiknya, ini biasanya berarti memperoleh persetujuan eksplisit atau memiliki mandat hukum yang jelas.

Pengontrol data harus memastikan:

- Keputusan tersebut dapat dijelaskan dan dapat diperdebatkan
- Data politik tidak pernah digunakan tanpa dasar hukum yang jelas
- Persetujuan diberikan secara bebas, spesifik, dan berdasarkan informasi yang memadai

Ras, Etnis, dan Keanggotaan Serikat Pekerja

Sistem AI sering memproses ras dan etnis baik secara langsung (misalnya, melalui formulir demografis) atau secara inferensial (misalnya, dari nama, bahasa, atau lokasi). Praktik ini sangat bermasalah dalam penyaringan resume, aplikasi pinjaman, platform perumahan, dan kepolisian serta pengawasan.

Demikian pula, keanggotaan serikat pekerja dapat disimpulkan dari isi email, aktivitas media sosial, atau perilaku organisasi, seringkali tanpa kesadaran atau persetujuan pengguna. Misalnya, sebuah organisasi menggunakan filter resume otomatis yang memprioritaskan kandidat dari perguruan tinggi kulit hitam karena pola perekrutan historis dalam kelompok pelatihan. Pemberi kerja tidak menyadari bias tersebut, tetapi dampak sistem yang tidak

merata terhadap minoritas etnis dapat merupakan diskriminasi tidak langsung dan pemrosesan data rasial yang melanggar hukum.

Bias Tersembunyi dalam Kumpulan Data Pelatihan AI

Banyak kumpulan data pelatihan mengandung keputusan historis yang bias, bahkan jika atribut sensitif telah dikecualikan. Teknik seperti pembelajaran yang sadar akan keadilan, evaluasi kontrafaktual, dan pengujian atribut sensitif sangat penting untuk mengidentifikasi dan mengurangi bias ini.

Dasar Hukum dan Persetujuan untuk Data Sensitif

Seperti yang telah dibahas sebelumnya dalam bab ini, Pasal 9 GDPR melarang pemrosesan data kategori khusus secara default. Pengontrol data harus mengidentifikasi salah satu dari beberapa pengecualian yang sempit untuk melanjutkan secara legal.

Pengecualian ini memerlukan justifikasi dan dokumentasi yang cermat. Tidak seperti data pribadi "biasa" (yang dapat diproses berdasarkan kepentingan sah, pelaksanaan kontrak, dan sebagainya), data sensitif membutuhkan dasar yang lebih kuat dan akuntabilitas yang lebih tinggi.

Persetujuan Eksplisit: Standar yang Tinggi

GDPR jelas mengenai persetujuan untuk pengumpulan dan penggunaan data sensitif. Persetujuan eksplisit harus:

- Diberikan secara bebas Spesifik
- Terinformasi
- Tidak ambigu
- Afirmatif (misalnya, tidak ada kotak yang sudah dicentang sebelumnya)
- Didokumentasikan dengan jelas

Dalam AI, persetujuan eksplisit sulit diperoleh dan bahkan lebih menantang untuk dipertahankan karena:

- ✓ Pengumpulan data tidak langsung atau inferensial
- ✓ Perubahan tujuan model dari waktu ke waktu
- ✓ Kompleksitas penjelasan yang diperlukan untuk membuat persetujuan bermakna

Misalnya, aplikasi kesehatan meminta persetujuan pengguna untuk menganalisis data detak jantung dan tidur. Persetujuan tersebut terkubur dalam kebijakan privasi yang panjang dan tidak menjelaskan penggunaan selanjutnya (misalnya, pemberian lisensi data kepada pihak ketiga untuk penelitian dan pengembangan deteksi emosi). Praktik ini tidak memenuhi standar persetujuan eksplisit dan berdasarkan informasi untuk pemrosesan data kesehatan.

Persetujuan vs. Kontrak dalam Sistem AI

Pengendali sering kali mengacaukan pelaksanaan kontrak (Pasal 6(1)(b) GDPR) dengan persetujuan (Pasal 9(2)(a)). Untuk data sensitif:

- ➢ Persetujuan harus eksplisit dan dapat dicabut

- Kontrak tidak cukup kecuali data sensitif tersebut benar-benar diperlukan untuk memenuhi kontrak
- Fitur kenyamanan atau "fitur tambahan yang diinginkan" tidak memenuhi syarat

Selalu nilai apakah fitur yang membutuhkan data sensitif tersebut penting, atau hanya opsional dan bergantung pada persetujuan.

Ketika Persetujuan Saja Tidak Cukup

Bahkan dengan persetujuan eksplisit, pemrosesan data sensitif dapat ilegal atau tidak etis jika:

- Subjek data berada dalam hubungan yang tidak seimbang kekuasaannya (misalnya, majikan-karyawan)
- Persetujuan tidak diberikan secara bebas (misalnya, membatasi akses layanan inti di balik persetujuan)
- Sistem AI menciptakan efek signifikan yang memicu perlindungan Pasal 22

Dalam sistem AI berisiko tinggi, regulator dapat mensyaratkan pengamanan tambahan, seperti DPIA, laporan transparansi, dan tinjauan berkelanjutan.

Pengamanan Teknis dan Organisasi

Sistem yang memproses data sensitif tanpa perlindungan yang kuat meningkatkan kemungkinan:

- Akses atau pelanggaran yang tidak sah
- Pencurian identitas
- Diskriminasi atau bias
- Sanksi peraturan
- Kerusakan reputasi

Pengontrol harus menerapkan langkah-langkah teknis dan organisasi (TOM) yang melampaui kontrol perlindungan data dasar. Untuk data sensitif, langkah-langkah ini harus mencerminkan peningkatan risiko yang terkait dengannya.

Langkah-Langkah Teknis Inti

Langkah-langkah utama yang digunakan untuk melindungi data sensitif meliputi:

- Enkripsi saat data disimpan dan saat data ditransmisikan (terutama untuk data biometrik dan kesehatan) Kontrol akses dan izin berbasis peran
- Anonimisasi atau pseudonimisasi jika memungkinkan
- Minimisasi data dan penyamaran tingkat bidang
- Lingkungan pelatihan model yang aman dengan pencatatan audit
- Alat penjelasan model untuk mendeteksi diskriminasi

Sebagai contoh, penyedia biometrik suara menerapkan privasi diferensial pada alur pelatihannya, memastikan bahwa tidak ada satu pun sidik suara yang dapat direkonstruksi dari keluaran model. Mereka juga menyimpan sidik suara di brankas terenkripsi yang dikontrol aksesnya dan merotasi kunci kriptografi setiap 90 hari. Langkah-langkah ini mengurangi dampak potensi kompromi.

Pembelajaran Terfederasi untuk Melindungi Data Kesehatan

Pembelajaran terfederasi memungkinkan pelatihan model terdesentralisasi; data pribadi tetap berada di perangkat pengguna, dan hanya pembaruan model anonim yang dibagikan. Pendekatan ini:

- Meminimalkan risiko sentralisasi data
- Membantu memenuhi persyaratan lokalisasi data
- Mendukung privasi di lingkungan yang sensitif (misalnya, rumah sakit, aplikasi kesehatan mental)

Catatan: Sistem terfederasi tetap memerlukan pengamanan agregasi, validasi, dan pengembalian yang aman.

Pengamanan Organisasi

Proses dan fungsi utama yang digunakan untuk melindungi data sensitif meliputi:

- DPIA (Penilaian Dampak Perlindungan Data) untuk semua pemrosesan data sensitif
- Kebijakan retensi dan penghapusan yang jelas
- Pelatihan bagi insinyur dan ilmuwan data tentang sensitivitas dan risiko
- Dewan peninjau atau komite etik untuk AI berisiko tinggi
- Pemeriksaan pemasok (misalnya, memastikan vendor biometrik pihak ketiga memenuhi standar GDPR)
- Prosedur dan pelatihan respons insiden dan pelanggaran yang formal
- Manajemen risiko formal dan penilaian risiko berkala

Organisasi juga harus memelihara jejak audit keputusan pemrosesan data sensitif, termasuk dasar hukum, rencana mitigasi risiko, dan pemantauan kinerja sistem.

Merancang untuk Minimisasi Data

Organisasi cenderung mengumpulkan data pribadi secara berlebihan "hanya untuk berjaga-jaga." Untuk kategori sensitif, ini berbahaya dan kemungkinan melanggar hukum.

Praktik terbaik untuk minimisasi data meliputi:

- ✓ Menggunakan data sintetis untuk pengujian awal
- ✓ Pengurangan fitur untuk mengecualikan atribut sensitif yang tidak diperlukan untuk akurasi model
- ✓ Melakukan penilaian proporsionalitas untuk menentukan apakah tujuan model membenarkan tingkat sensitivitas input
- ✓ Melakukan audit dataset berkala untuk mendeteksi dan menghapus data yang tidak perlu atau sensitif, termasuk data yang mungkin disimpulkan selama pemrosesan

Pertimbangan Khusus dalam Aplikasi AI

Pengumpulan dan pemrosesan data sensitif memerlukan pertimbangan tambahan.

Bias dan Keadilan dalam Atribut Sensitif

Jenis data sensitif sering berkorelasi dengan kelompok yang secara historis terpinggirkan atau dilindungi. Ketika model AI menggunakan atau menyimpulkan atribut seperti ras, agama, jenis kelamin, atau status kesehatan—bahkan secara tidak langsung—

mereka dapat mereplikasi atau memperburuk diskriminasi yang ada. Kesalahan umum meliputi:

- Bias data pelatihan (misalnya, keputusan perekrutan historis yang mengecualikan kelompok tertentu)
- Diskriminasi proksi (misalnya, kode pos sebagai pengganti ras)
- Overfitting pada fitur sensitif yang memengaruhi prediksi
- Pengecualian yang tidak beralasan terhadap kelompok sensitif dari cakupan model

Penilaian keadilan harus mencakup evaluasi kuantitatif tentang bagaimana model berkinerja di berbagai kelompok demografis dan tinjauan kualitatif tentang konteks sosial.

Contoh sistem AI persetujuan pinjaman yang dilatih pada data pinjaman historis menunjukkan tingkat penolakan yang jauh lebih tinggi untuk pelamar yang diduga berasal dari latar belakang minoritas tertentu, meskipun ras bukan variabel masukan. Korelasi tersebut dikodekan dalam fitur kode pos dan pendapatan, yang memicu kekhawatiran tentang diskriminasi tidak langsung.

Pertanyaan Audit untuk Keadilan Atribut Sensitif

Berikut beberapa pertanyaan yang dapat digunakan untuk memastikan model AI bebas dari bias:

- Apakah ada atribut sensitif yang secara eksplisit digunakan dalam model?
- Dapatkah fitur non-sensitif berfungsi sebagai proksi?
- Bagaimana akurasi model bervariasi di berbagai kelompok yang dilindungi?
- Apakah ada kelompok demografis yang kurang terwakili dalam data pelatihan?
- Apakah batasan keadilan diterapkan dan didokumentasikan?
- Apakah keadilan model tetap berlaku dari waktu ke waktu saat data baru diperkenalkan?

Audit rutin, terutama setelah pembaruan model, sangat penting untuk AI yang melibatkan kategori sensitif.

Inferensi Data Sensitif dari Input Non-Sensitif

Sistem AI mampu melakukan pelanggaran privasi inferensial, di mana model menyimpulkan karakteristik sensitif dari data yang tidak terkait. Contohnya meliputi:

- Afiliasi keagamaan dari pola bahasa
- Orientasi seksual dari koneksi grafik sosial
- Kondisi kesehatan dari data sensor yang dapat dikenakan
- Kecenderungan politik dari riwayat penelusuran

Inferensi ini dapat memicu perlindungan Pasal 9 GDPR meskipun atribut tersebut tidak dikumpulkan secara eksplisit.

Contoh aplikasi kencan menggunakan LLM untuk menyarankan kecocokan berdasarkan riwayat obrolan dan pola interaksi. Seiring waktu, aplikasi tersebut mulai menyimpulkan kondisi kesehatan mental dan volatilitas emosional pengguna. Tidak ada data kesehatan

mental yang dikumpulkan; namun, inferensi dapat mengakibatkan hasil yang diskriminatif atau manipulasi profil, yang berpotensi melanggar aturan data sensitif.

Sensitivitas Melalui Proksi: Hal yang Perlu Diwaspadai

Saat melatih dan menggunakan sistem AI, organisasi harus mewaspadai:

- ✓ Korelasi tinggi antara fitur dan kategori yang dilindungi
- ✓ Model yang dilatih pada dataset terbuka yang tidak seimbang atau bias
- ✓ Atribut sensitif yang tidak berlabel yang disimpulkan selama pengujian atau penerapan
- ✓ Vendor yang menawarkan layanan "pengayaan demografis"
- ✓ Pergeseran pasca-penerapan yang mengubah perilaku model atau keadilan di seluruh kelompok demografis

Pengontrol bertanggung jawab atas inferensi hilir—bahkan ketika tidak dirancang secara eksplisit.

Penyalahgunaan Tujuan dan Penggunaan Sekunder

Data sensitif yang dikumpulkan untuk satu tujuan (misalnya, pelacakan kesehatan) tidak boleh digunakan kembali untuk tujuan baru yang tidak kompatibel (misalnya, prediksi perilaku untuk penetapan harga asuransi) kecuali:

- ➢ Subjek data secara eksplisit menyetujui
- ➢ Penggunaan kembali tersebut termasuk dalam pengecualian hukum yang jelas
- ➢ Tujuan baru tersebut terbukti kompatibel dengan tujuan aslinya

Sistem AI seringkali berkembang pesat, dan organisasi cenderung menggunakan sistem AI yang ada untuk tujuan baru, sehingga penyalahgunaan tujuan menjadi risiko yang berkelanjutan.

Perusahaan yang sama kemudian menggunakan kembali model pengenalan emosi yang dikembangkan untuk analitik game untuk mengevaluasi kandidat pekerjaan selama wawancara. Kumpulan data asli berisi ekspresi wajah yang diambil dalam konteks non-profesional, menimbulkan pertanyaan tentang batasan tujuan dan keadilan.

DPIA dan Manajemen Risiko

DPIA wajib dilakukan berdasarkan Pasal 35 GDPR ketika pemrosesan data "kemungkinan besar akan menimbulkan risiko tinggi terhadap hak dan kebebasan individu." Hal ini hampir selalu terjadi pada sistem AI yang memproses kategori data khusus.

Pemicu risiko tinggi meliputi:

- ❖ Pemantauan sistematis terhadap data sensitif
- ❖ Pengambilan keputusan otomatis dengan dampak hukum atau dampak signifikan serupa
- ❖ Pemrosesan data populasi rentan
- ❖ Penggunaan data biometrik atau kesehatan dalam skala besar

Pengendali data harus mendokumentasikan dan membenarkan keputusan mereka, baik mereka melakukan DPIA atau tidak.

Apa yang Harus Dimasukkan dalam DPIA untuk AI Sensitif

DPIA yang kuat untuk AI yang melibatkan data sensitif harus mencakup:

- Deskripsi sistem AI, termasuk input, logika, dan output
- Tujuan pemrosesan dan justifikasi sensitivitas data
- Penilaian risiko potensi bahaya bagi individu
- Evaluasi kemungkinan dan tingkat keparahan risiko
- Pengamanan, langkah-langkah teknis, dan kontrol kepatuhan
- Masukan pemangku kepentingan, termasuk dari DPO, hukum, dan komunitas yang terdampak

Misalnya, sebuah universitas memperkenalkan AI deteksi emosi berbasis suara untuk memantau kesejahteraan mental mahasiswa. DPIA mengidentifikasi risiko kesalahan klasifikasi, penyalahgunaan data, dan stigmatisasi. Mitigasi mencakup persetujuan, pemrosesan lokal, penghapusan data mentah, dan komunikasi yang kuat dengan mahasiswa.

Pertanyaan-pertanyaan Penting untuk Proyek AI Sensitif
▪ Apakah sistem tersebut menyimpulkan atribut sensitif? ▪ Apakah individu menyadari bahwa data sensitif mereka sedang digunakan? ▪ Dapatkah pengguna memilih untuk tidak ikut serta tanpa kehilangan akses ke layanan inti? ▪ Apakah ada mitra atau vendor hilir yang menggunakan data yang sama? ▪ Apakah risiko telah ditinjau oleh tim hukum, etika, dan keamanan?
Jika jawaban atas salah satu pertanyaan ini tidak jelas atau negatif, maka diperlukan DPIA—dan berpotensi konsultasi dengan regulator.

Peran Register Risiko dan Dewan Peninjau

Sistem AI berisiko tinggi yang melibatkan data sensitif harus dikatalogkan dalam register risiko AI internal, termasuk:

- Deskripsi sistem dan pemilik
- Tingkat keparahan risiko berdasarkan dampak dan kemungkinan
- Tanggal DPIA dan penilaian ulang
- Paparan regulasi
- Riwayat insiden (jika ada)
- Status mitigasi
- Tautan ke dokumentasi pendukung untuk kesiapan audit

Beberapa organisasi juga membentuk dewan peninjau AI atau panel etika untuk menilai penerapan yang sangat sensitif atau kontroversial, praktik terbaik tata kelola yang mendukung kepatuhan dan kepercayaan publik.

Hak Subjek Data dan Data Sensitif

Pemrosesan data sensitif oleh AI meningkatkan pentingnya dan kompleksitas hak subjek data, sehingga memerlukan transparansi yang lebih kuat, opsi penolakan, dan perlindungan terhadap keputusan otomatis.

Ekspektasi Hak yang Lebih Tinggi

Ketika sistem AI memproses kategori data khusus, subjek data berhak atas transparansi, kontrol, dan ganti rugi yang lebih tinggi. Semua hak inti berdasarkan undang-undang perlindungan data tetap berlaku, tetapi penerapannya dapat menjadi lebih kompleks dan kritis.

Hak-hak utama subjek data harus mencakup:

- **Hak untuk diinformasikan:** Mengapa data sensitif diproses? Bagaimana data tersebut diperoleh?
- **Hak akses:** Data sensitif apa yang disimpan, dan bagaimana data tersebut digunakan?
- **Hak untuk perbaikan:** Dapatkah pengguna memperbaiki atribut sensitif yang disimpulkan atau salah klasifikasi?
- **Hak untuk penghapusan:** Dapatkah pengguna menuntut penghapusan catatan biometrik atau kesehatan? (Ada tantangan khusus terkait penghapusan catatan pelatihan individu dari sistem AI yang dibahas di seluruh buku ini.)
- **Hak untuk keberatan:** Dapatkah pengguna menolak pembuatan profil atau pemrosesan otomatis berdasarkan data sensitif?
- **Hak untuk tidak tunduk pada keputusan otomatis:** Dengan efek hukum atau efek signifikan serupa (GDPR Pasal 22, CPRA 1798.121).

Data Sensitif Menambah Risiko pada Permintaan Hak AI

Pemrosesan data sensitif seringkali membuat permintaan subjek data (DSR) menjadi lebih mendesak, lebih intensif sumber daya, dan lebih mungkin menyebabkan keluhan jika salah penanganan.

Tantangan meliputi:

- ✓ Memberikan penjelasan yang mudah dipahami ketika keputusan dibuat oleh model yang tidak transparan
- ✓ Mengoreksi data sensitif yang disimpulkan yang tidak pernah diberikan pengguna secara eksplisit
- ✓ Mempertahankan kinerja model sambil menghapus data pelatihan yang terkait dengan pengguna tertentu yang ingin memilih keluar
- ✓ Menavigasi konflik antara permintaan penghapusan dan persyaratan retensi peraturan, kontrak, atau penelitian

Misalnya, seorang pencari kerja meminta penghapusan data yang digunakan oleh alat perekrutan AI setelah ditolak. Perusahaan harus menentukan apakah data tersebut terkait dengan bobot model, apakah pelatihan ulang layak dilakukan, dan apakah penghapusan data akan berdampak pada kepatuhan terhadap persyaratan dokumentasi anti-diskriminasi.

Cara Menanggapi Permintaan Hak Data (DSR) yang Melibatkan Data Sensitif
➢ Konfirmasi identitas dengan sangat hati-hati (terutama untuk data biometrik) ➢ Berikan penjelasan yang jelas tentang apakah data dikumpulkan atau disimpulkan ➢ Identifikasi keputusan atau konsekuensi apa pun yang terkait dengan data ➢ Jelaskan dasar hukum untuk pemrosesan dan pihak ketiga mana pun yang terlibat ➢ Catat semua tindakan yang diambil dan jangka waktu yang dipenuhi
Pengontrol harus menyiapkan prosedur operasi standar (SOP) untuk permintaan hak yang melibatkan kategori informasi sensitif.

Hak dan Batasan Opt-out dalam Konteks AI

Berdasarkan GDPR dan undang-undang seperti CPRA, individu memiliki hak untuk menolak jenis pemrosesan data tertentu, terutama ketika melibatkan pembuatan profil, iklan bertarget, atau pengambilan keputusan otomatis.

Dalam sistem AI, mekanisme opt-out menjadi lebih penting dan lebih kompleks. Pengguna harus dapat menolak pemrosesan untuk tujuan sekunder, menolak pembuatan profil, atau menolak untuk tunduk pada keputusan otomatis, terutama ketika keputusan tersebut dapat berdampak signifikan pada mereka (misalnya, dalam konteks perekrutan, pinjaman, atau perawatan kesehatan).

Mekanisme opt-out yang sesuai harus:

- ❖ Disajikan dengan jelas dan mudah digunakan
- ❖ Terperinci, memungkinkan individu untuk menolak penggunaan spesifik terkait AI (misalnya, iklan bertarget, penyaringan otomatis)
- ❖ Efektif, artinya organisasi harus menindaklanjuti permintaan tersebut dan mengkonfirmasi penghentian pemrosesan

Misalnya, papan lowongan kerja menggunakan model AI untuk memberi peringkat pelamar berdasarkan kesesuaian budaya yang disimpulkan. Setelah memilih untuk tidak ikut serta, satu kandidat masih dinilai oleh sistem dan disaring, yang merupakan pelanggaran hak untuk keberatan berdasarkan Pasal 21 GDPR dan dapat memerlukan pelaporan regulasi jika tidak terselesaikan. Hak untuk tidak ikut serta bukanlah hak mutlak. Berdasarkan GDPR, hak tersebut dapat diabaikan jika:

- Pemrosesan diperlukan untuk kontrak atau kewajiban hukum
- Terdapat kepentingan yang sah dan mendesak yang mengesampingkan hak individu

Namun, ketika pembuatan profil digunakan untuk pemasaran langsung, hak untuk tidak ikut serta adalah mutlak dan harus dihormati tanpa pengecualian.

Pengontrol harus mendokumentasikan dan menegakkan pilihan untuk tidak ikut serta di semua sistem dan vendor, termasuk yang terkait dengan pelatihan model dan konteks pembuatan profil.

Hak dalam Pengambilan Keputusan Otomatis Berisiko Tinggi

Ketika sistem AI memproses data sensitif dan secara signifikan memengaruhi individu (misalnya, penolakan dari layanan, pekerjaan, atau kredit), pembatasan pengambilan keputusan otomatis berlaku.

Berdasarkan Pasal 22 GDPR:

- Individu memiliki hak untuk tidak tunduk pada keputusan yang sepenuhnya otomatis.
- Pengecualian (misalnya, persetujuan eksplisit atau kebutuhan kontraktual) harus mencakup perlindungan seperti:
    - Keterlibatan manusia yang bermakna
    - Hak untuk menentang keputusan
    - Penjelasan yang transparan tentang logika dan konsekuensi

Sistem AI yang menggunakan data sensitif untuk mendorong keputusan kelayakan atau penetapan harga (misalnya, premi asuransi berbasis kesehatan, penolakan akses biometrik) sangat mungkin berada di bawah perlindungan ini.

4.6 RINGKASAN

Sistem AI menghadirkan tantangan unik terhadap prinsip-prinsip perlindungan data yang telah lama ada. Seiring organisasi memanfaatkan AI untuk menganalisis, memprediksi, dan membuat keputusan tentang individu, konsep privasi mendasar seperti pemberitahuan, pilihan, persetujuan, dan pembatasan tujuan harus ditafsirkan ulang dengan cara yang mengatasi kompleksitas dan ketidakjelasan algoritma modern. Individu harus diberi tahu tentang bagaimana data mereka digunakan, terutama ketika pengambilan keputusan otomatis terlibat. Namun, pemberitahuan privasi tradisional seringkali kurang memadai untuk menjelaskan logika atau implikasi pembelajaran mesin.

Demikian pula, persetujuan harus bermakna, terinformasi, dan eksplisit—suatu hal yang sulit ketika sistem AI beroperasi di balik layar atau bergantung pada data yang disimpulkan. Pembatasan tujuan sering diuji ketika model yang dilatih untuk satu fungsi kemudian digunakan kembali untuk fungsi lain, meningkatkan risiko penggunaan sekunder yang tidak sah. Untuk mengurangi risiko ini, prinsip minimalisasi data dan privasi berdasarkan desain harus diintegrasikan di seluruh siklus hidup AI, didukung oleh teknik seperti pembelajaran federasi, privasi diferensial, dan arsitektur model modular.

Secara paralel, pengendali AI (mereka yang menentukan tujuan dan cara pemrosesan) memiliki kewajiban yang luas berdasarkan undang-undang seperti GDPR. Mereka harus melakukan DPIA (Penilaian Dampak Perlindungan Data) untuk pemrosesan berisiko tinggi, terutama ketika keputusan otomatis memengaruhi hak-hak orang. Saat melibatkan pemroses pihak ketiga, pengendali harus menetapkan kontrak yang kuat dan memastikan pengawasan berkelanjutan untuk menghindari "penyimpangan pemroses".

Banyak ekosistem AI bersifat global, memicu aturan tentang transfer data lintas batas dan memerlukan mekanisme seperti Klausul Kontrak Standar atau Penilaian Dampak Transfer. Pengendali juga bertanggung jawab untuk menghormati hak subjek data, termasuk akses,

penghapusan, dan keberatan terhadap pembuatan profil, bahkan ketika modelnya kompleks dan tidak transparan.

Manajemen insiden menjadi lebih bernuansa dengan penggunaan AI, terutama ketika pelanggaran melibatkan inversi model, inferensi keanggotaan, atau data pelatihan yang sensitif. Untuk memenuhi persyaratan akuntabilitas, organisasi harus memelihara dokumentasi terperinci, termasuk kartu model, peta data, register risiko, dan log audit.

Area fokus yang penting adalah penggunaan kategori data sensitif atau khusus dalam AI, termasuk biometrik, data kesehatan, opini politik, kepercayaan agama, dan afiliasi ras atau serikat pekerja. Jenis data ini memerlukan perlindungan yang lebih tinggi karena potensinya untuk menyebabkan kerugian, memfasilitasi diskriminasi, atau memungkinkan pengawasan. Pengontrol data harus mengidentifikasi dasar hukum untuk pemrosesan, seringkali persetujuan eksplisit, dan menerapkan perlindungan tambahan. Sistem AI sering kali menyimpulkan atribut sensitif bahkan ketika atribut tersebut tidak dikumpulkan secara eksplisit, sebuah risiko yang mempersulit kepatuhan dan menimbulkan kekhawatiran etis. Penilaian Dampak Perlindungan Data (DPIA) seringkali wajib dalam skenario ini, dan perlindungan harus melampaui enkripsi dan kontrol akses untuk mencakup mitigasi bias, pengujian keadilan, dan transparansi pengguna yang kuat. Pengontrol data juga harus memastikan bahwa individu dapat menggunakan hak mereka terkait data sensitif, terutama ketika keputusan otomatis dapat secara signifikan memengaruhi kehidupan mereka. Seiring dengan meningkatnya kemampuan AI, demikian pula tanggung jawab untuk menjunjung tinggi martabat, otonomi, dan privasi mereka yang datanya mendukung sistem ini.

BAB 5
HUKUM SEKTORAL DAN PERDATA

5.1 HUKUM KEKAYAAN INTELEKTUAL DAN AI

Seiring sistem AI menjadi lebih mampu menyerap, menafsirkan, dan mereplikasi ekspresi manusia, pertanyaan seputar kekayaan intelektual (IP) telah menjadi pusat perhatian. Inti dari perdebatan ini adalah ketegangan mendasar: sistem AI membutuhkan volume data yang besar untuk belajar secara efektif, namun sebagian besar data tersebut mungkin dilindungi oleh hak cipta, paten, rahasia dagang, atau hak kekayaan intelektual lainnya. Bagian ini mengeksplorasi bagaimana hukum kekayaan intelektual berlaku untuk AI, khususnya dalam konteks pengumpulan data dan pelatihan model, dan memberikan panduan praktis untuk menavigasi lanskap hukum yang berkembang pesat.

Dasar-Dasar Kekayaan Intelektual dan AI

Rezim hukum kekayaan intelektual mencakup berbagai perlindungan yang dimaksudkan untuk mendorong kreativitas, inovasi, dan pengungkapan. Karena sistem AI semakin banyak mengumpulkan konten yang dilindungi kekayaan intelektual, doktrin perlindungan hukum yang ada sedang diuji dengan cara baru. Bagian ini menjelaskan bagaimana kategori kekayaan intelektual tradisional —hak cipta, paten, rahasia dagang, dan hak basis data—berlaku untuk pengembangan dan penggunaan AI.

Hukum Hak Cipta: Pusat Kontroversi Pelatihan AI

Hukum hak cipta melindungi “karya asli ciptaan” yang terabadikan dalam media yang berwujud, termasuk teks, musik, gambar, video, dan kode perangkat lunak. Bagi sistem AI generatif yang bergantung pada pengambilan konten publik dari web, hal ini menghadirkan ladang ranjau hukum dan etika. Para kreator konten manusia menyuarakan kekhawatiran tentang sistem AI yang mengancam untuk menghilangkan peran perantara kreator, merampas kemampuan mereka untuk mencari nafkah melalui ekspresi kreatif.

Pertanyaan hukum utama meliputi:

- Apakah penggunaan karya berhak cipta untuk pelatihan AI merupakan pelanggaran hak reproduksi?
- Apakah pelatihan tersebut termasuk penggunaan "transformatif" berdasarkan doktrin penggunaan wajar?
- Apakah keluaran yang dihasilkan AI melanggar hak cipta jika sangat mirip dengan karya berhak cipta?

Pertanyaan-pertanyaan ini sedang diuji di pengadilan di seluruh dunia, seringkali tanpa preseden yang jelas.

Pada tahun 2023, Getty Images menggugat Stability AI karena diduga menggunakan jutaan gambar berhak cipta untuk melatih model generatifnya tanpa izin atau kompensasi. Klaim Getty bergantung pada pelanggaran hak cipta dan merek dagang, dengan alasan bahwa penggunaan gambar-gambarnya, bahkan untuk tujuan pelatihan, bukanlah "transformatif"

dan secara langsung bersaing dengan bisnis lisensinya. (Kasus ini masih dalam proses hingga saat penulisan buku ini.)

Dasar-Dasar Penggunaan Wajar

Undang-Undang Hak Cipta AS mengizinkan penggunaan materi berhak cipta tertentu yang tidak sah berdasarkan doktrin penggunaan wajar. Pengadilan mempertimbangkan empat faktor:

- Tujuan dan karakter penggunaan (misalnya, komersial vs. nirlaba, transformatif vs. duplikatif)
- Sifat karya berhak cipta
- Jumlah dan substansi bagian yang digunakan
- Pengaruh terhadap pasar untuk karya berhak cipta asli

Pengembang AI sering mengklaim bahwa pelatihan merupakan penggunaan wajar karena model tidak menyalin atau mereplikasi karya persis tetapi belajar dari menganalisis pola dan hubungan. Namun, argumen ini semakin sering ditantang di pengadilan.

Hukum Paten dan Invensi yang Dihasilkan AI

Hukum paten melindungi invensi baru, bermanfaat, dan tidak jelas. Penemu yang ingin memonetisasi invensi mereka dapat memilih untuk mempublikasikan desain invensi mereka ke Kantor Paten AS, yang kemudian memberikan hak eksklusif kepada penemu untuk menjual atau melisensikan invensi tersebut untuk jangka waktu tertentu (negara lain beroperasi serupa). Meskipun fokus wacana AI/IP biasanya pada hak cipta dan data, masalah paten muncul dalam dua cara:

- ✓ Penggunaan data atau algoritma yang dipatenkan dalam pelatihan
- ✓ Kemungkinan dipatenkannya invensi yang dihasilkan AI

Yang terakhir telah menarik perhatian yang signifikan. Haruskah invensi yang dirancang oleh sistem AI (tanpa penemu manusia) dapat dipatenkan? Demikian pula, haruskah perintah tersebut digunakan untuk menciptakan karya yang dapat dipatenkan yang juga dapat dipatenkan atau dilindungi oleh hak cipta?

Sistem AI DABUS disebut sebagai satu-satunya penemu dalam aplikasi paten yang diajukan di lebih dari selusin yurisdiksi. Meskipun Afrika Selatan memberikan paten tersebut, sebagian besar kantor paten (termasuk USPTO dan EPO) menolak permohonan tersebut, dengan alasan bahwa menurut hukum yang berlaku saat ini hanya orang perseorangan yang dapat menjadi penemu.

Bisakah AI Menjadi Penemu?

Secara global, pengadilan dan kantor paten telah menolak gagasan bahwa AI dapat memenuhi syarat sebagai penemu. Pengadilan Sirkuit Federal AS dalam kasus Thaler v. Vidal (2022) memutuskan bahwa Undang-Undang Paten mensyaratkan adanya penemu manusia, yang secara efektif menghentikan aplikasi bergaya DABUS untuk saat ini.

Rahasia Dagang dan Model AI

Rahasia dagang adalah bentuk kekayaan intelektual, biasanya berupa desain, metode, rumus, data, dan pengetahuan tentang suatu produk atau proses, yang pemiliknya memilih untuk menyembunyikannya dari publik, dan kerahasiaannya memberikan nilai ekonomi. Hukum rahasia dagang melindungi informasi bisnis rahasia yang memperoleh nilai ekonomi karena tidak diketahui secara umum. Dalam konteks AI, contoh rahasia dagang meliputi:

- Dataset pelatihan milik perusahaan
- Arsitektur model dan hiperparameter
- Metodologi penyempurnaan
- Algoritma optimasi inferensi
- Teknik pra-pemrosesan dan augmentasi data
- Metrik dan tolok ukur evaluasi
- Infrastruktur penyebaran dan metode penskalaan
- Teknik augmentasi data

Karena hak cipta tidak melindungi fakta atau ide, dan paten memerlukan pengungkapan publik, banyak perusahaan mengandalkan perlindungan rahasia dagang untuk keunggulan kompetitif AI.

Tantangan terkait perlindungan rahasia dagang meliputi:

- Memastikan kerahasiaan dataset dalam lingkungan kolaboratif
- Mencegah rekayasa balik output yang dapat mengungkapkan data pelatihan dan/atau algoritma
- Menyeimbangkan transparansi untuk tinjauan etis dengan pelestarian rahasia dagang
- Paparan rantai pasokan dari penyedia layanan pihak ketiga yang menangani dataset atau model sensitif

Meskipun versi awal GPT OpenAI dipublikasikan secara terbuka, versi selanjutnya seperti GPT-4 menjadi hak milik. OpenAI menyebutkan keamanan dan daya saing sebagai alasan untuk menahan data pelatihan dan detail model. Strategi rahasia dagang ini juga melindungi perusahaan dari klaim (sampai batas tertentu) oleh organisasi yang mencurigai pelanggaran IP.

Hak Basis Data dan Variabilitas Regional

Di Uni Eropa dan yurisdiksi tertentu, hak basis data sui generis melindungi investasi besar yang dilakukan dalam pengumpulan data. Hak-hak ini dapat membatasi ekstraksi dan penggunaan kembali data yang tidak sah, bahkan jika titik data individual itu sendiri tidak dilindungi oleh hak cipta.

Hal ini menjadi relevan ketika model AI dilatih pada kumpulan data terstruktur seperti catatan medis, basis data ilmiah, atau arsip keuangan. Di Uni Eropa dan tempat lain, pemilik basis data ini dapat melarang penggunaannya dalam pelatihan sistem AI.

Direktif Basis Data Uni Eropa

Berdasarkan Direktif Basis Data Uni Eropa (Direktif 96/9/EC), pemegang hak dapat melarang pengambilan atau penggunaan kembali bagian-bagian substansial dari suatu basis data. Pengembang AI mungkin perlu memperoleh lisensi, bahkan jika konten tersebut tidak dilindungi hak cipta.

Penggunaan Karya yang Dilindungi dalam Pelatihan AI

Meskipun hukum kekayaan intelektual (KI) dasar menetapkan batasan tentang apa yang dapat dan tidak dapat digunakan, penerapannya pada praktik pelatihan AI tidaklah mudah. Bagian ini membahas bagaimana hukum KI membatasi (atau mengizinkan) penggunaan karya yang dilindungi dalam kumpulan data pelatihan AI, dengan fokus pada litigasi saat ini, pendekatan perizinan, dan tanggung jawab pengembang.

Dilema Data Pelatihan

Melatih sistem AI seringkali membutuhkan kumpulan data yang sangat besar, biasanya diambil dari situs web, forum, artikel berita, repositori seni, dan basis kode. Banyak dari sumber-sumber ini berisi konten yang dilindungi hak cipta atau konten yang diatur oleh lisensi yang membatasi.

Misalnya, GitHub Copilot, yang awalnya didukung oleh Codex OpenAI, menyarankan penyelesaian kode berdasarkan kode dari repositori GitHub publik. Pada tahun 2022, gugatan class-action menuduh bahwa Copilot mereproduksi kode sumber terbuka tanpa atribusi yang tepat, yang melanggar lisensi sumber terbuka seperti Lisensi Publik Umum GNU (GPL). Gugatan ini menimbulkan pertanyaan-pertanyaan sulit:

- Apakah pelatihan pada kode sumber terbuka memerlukan kepatuhan terhadap ketentuan lisensinya?
- Apakah penyelesaian kode merupakan karya turunan?
- Haruskah pengembang bertanggung jawab atas penggunaan kode yang disarankan oleh Copilot?

Model Lisensi untuk Data Pelatihan

Untuk menghindari sengketa kekayaan intelektual, beberapa pengembang AI beralih ke sumber data yang dilisensikan secara eksplisit atau data sintetis. Pendekatan lisensi meliputi:

- Konten domain publik (misalnya, data pemerintah AS, Project Gutenberg)
- Lisensi *Creative Commons* (CC), yang mungkin mengizinkan penggunaan kembali dengan syarat tertentu
- Perjanjian lisensi berbayar (misalnya, Shutterstock melisensikan perpustakaan gambarnya untuk pelatihan)
- Pembuatan data sintetis, yang mengurangi ketergantungan pada kumpulan data dunia nyata

Praktik Terbaik: Log Asal Usul Data

Memelihara catatan sumber data dan ketentuan lisensi semakin penting. Pengembang AI mengadopsi log asal usul data untuk melacak asal, hak penggunaan, dan persyaratan atribusi untuk setiap kumpulan data, sehingga memungkinkan posisi yang dapat dipertahankan jika terjadi klaim kekayaan intelektual. Asal usul data dibahas secara rinci di Bab 9.

Output AI dan Masalah Pelanggaran

Meskipun sebagian besar percakapan tentang kekayaan intelektual (KI) seputar AI berfokus pada data pelatihan, perhatian hukum semakin bergeser ke apa yang dihasilkan sistem AI. Karena model generatif menciptakan teks, gambar, video, kode, dan musik, garis antara inspirasi dan pelanggaran menjadi sulit untuk ditarik. Bagian ini mengeksplorasi apakah output yang dihasilkan AI dapat melanggar KI yang ada, dan siapa yang bertanggung jawab ketika hal itu terjadi.

Kapan Output AI Melanggar Hak Cipta?

Apakah output AI melanggar hak cipta biasanya bergantung pada dua faktor. Pertama, apakah output tersebut mirip dengan karya yang dilindungi (dan sejauh mana), dan kedua, akses apa yang dimiliki sistem AI terhadap karya asli selama pelatihan?

Jika model generatif mereproduksi materi berhak cipta, baik itu paragraf kata demi kata, foto yang hampir identik, atau melodi yang tidak dapat dibedakan dari lagu yang sudah dikenal, hal itu mungkin melanggar hak cipta, bahkan jika pengembang tidak bermaksud demikian.

Alat seperti Suno dan Udio memungkinkan pengguna untuk menghasilkan musik dalam berbagai genre. Para kritikus khawatir bahwa output dapat meniru lagu-lagu yang sudah ada, yang menyebabkan pelanggaran serupa dengan gugatan *Blurred Lines*, di mana keluarga Marvin Gaye dan Bridgeport Music berpendapat bahwa lagu tersebut melanggar hak cipta lagu *Gaye, Got to Give It Up*. Dalam kasus itu, juri menemukan pelanggaran hak cipta berdasarkan nuansa dan gaya (bukan peniruan langsung), yang menimbulkan kekhawatiran terhadap output musik AI yang meniru lagu-lagu ikonik.

Dalam contoh lain, Denmark telah mengusulkan undang-undang untuk memberikan perlindungan hak cipta kepada warganya atas wajah, suara, dan tubuh mereka, sebagai solusi hukum potensial terkait pembuatan deepfake.

Atribusi Output dan Akuntabilitas Hukum

Masalah yang lebih rumit adalah menentukan siapa yang bertanggung jawab jika output yang dihasilkan AI melanggar hak kekayaan intelektual. Opsi yang mungkin termasuk:

- Pengembang model
- Entitas yang melatihnya dengan data yang melanggar hak cipta
- Pengguna akhir yang memicu dan menggunakan output
- Model itu sendiri (meskipun tidak diakui secara hukum sebagai agen atau orang di sebagian besar yurisdiksi)

Tanggung jawab cenderung jatuh pada pihak manusia atau korporasi berdasarkan teori pelanggaran tidak langsung atau tanggung jawab kontributif.

Kasus Penting: Andersen v. Stability AI (2023)

Dalam gugatan ini, sekelompok seniman menuduh bahwa model gambar AI seperti Stable Diffusion menghasilkan karya turunan yang meniru gaya seni mereka. Penggugat berpendapat bahwa hanya dengan memberikan perintah "dengan gaya [nama seniman]" sudah menghasilkan konten yang melanggar hak cipta. Pengadilan belum memberikan putusan yang pasti, tetapi kasus ini menunjukkan bagaimana atribusi hasil karya dan identitas artistik saling berkaitan.

Kepemilikan dan Perlindungan Konten yang Dihasilkan AI

Selain risiko pelanggaran, organisasi juga harus menentukan apakah dan bagaimana output yang dihasilkan AI dapat dilindungi berdasarkan hukum kekayaan intelektual yang ada. Ini termasuk pertanyaan tentang hak cipta, kepengarangan, dan kemampuan untuk mengkomersialkan output. Tidak banyak preseden hukum dalam hal ini.

Hak Cipta Karya yang Dihasilkan AI

Berdasarkan hukum hak cipta AS dan sebagian besar hukum hak cipta internasional, hanya karya yang dibuat oleh penulis manusia yang memenuhi syarat untuk perlindungan. Kantor Hak Cipta AS telah menegaskan kembali prinsip ini dalam beberapa pernyataan panduan dan kasus pengadilan.

Pada tahun 2023, Kantor Hak Cipta AS mencabut perlindungan untuk gambar dalam novel grafis *Zarya of the Dawn* yang dibuat dengan alat AI Midjourney, dengan alasan bahwa gambar-gambar tersebut tidak memiliki kepengarangan manusia. Narasi tekstual, yang ditulis oleh manusia, tetap dilindungi, tetapi gambar-gambar tersebut tidak. Kasus ini menggambarkan kebijakan "kepengarangan manusia" Kantor Hak Cipta.

Kepengarangan Bersama dan AI sebagai Alat

Misalkan AI digunakan sebagai alat, bukan sebagai agen kreatif. Dalam kasus seperti itu, pengadilan mungkin masih memberikan perlindungan, asalkan manusia menjalankan kendali dan penilaian kreatif yang cukup, sehingga beberapa pihak membedakan antara:

- **Penggunaan berbasis alat:** Manusia memandu dan mengedit output (memenuhi syarat hak cipta)
- **Generasi otonom:** Model menciptakan dengan masukan manusia minimal (tidak memenuhi syarat)

Dalam konteks komersial, perbedaan ini penting ketika mendaftarkan hak cipta, melisensikan media yang dihasilkan AI, atau menegaskan eksklusivitas atas konten bermerek.

Strategi Mitigasi Risiko dan Kepatuhan

Mengingat ketidakpastian mengenai risiko dan ketidakpastian yang berkembang seputar IP dan AI, organisasi harus mengadopsi strategi kepatuhan proaktif. Bagian ini mencakup taktik tata kelola utama untuk mengurangi paparan dan membangun praktik yang dapat dipertahankan.

Membangun Kerangka Tata Kelola AI yang Sadar Kekayaan Intelektual

Kerangka tata kelola perusahaan yang menangani kekayaan intelektual harus mencakup:

- ✓ Uji tuntas untuk sumber data pelatihan, untuk memastikan organisasi tidak akan dikenai tindakan hukum oleh pemilik sumber data
- ✓ Tinjauan kewajiban lisensi (terutama sumber terbuka atau Creative Commons) untuk data pelatihan
- ✓ Kontrol atas penggunaan keluaran model yang mungkin memerlukan tinjauan keluaran untuk kecukupan hukum
- ✓ Pencatatan asal usul data dan silsilah model
- ✓ Kebijakan untuk mengkodifikasi aturan bisnis untuk hal-hal di atas

Kegiatan-kegiatan ini harus tertanam dalam proses siklus hidup AI, khususnya tahap desain dan pengembangan model, penerapan, dan audit.

Contoh Daftar Periksa: Tinjauan Risiko Kekayaan Intelektual AI

- ➢ Apakah semua dataset pelatihan diperoleh atau dilisensikan secara legal?
- ➢ Apakah lisensi sumber terbuka dihormati (misalnya, GPL, MIT, CC-BY)?
- ➢ Apakah model menghasilkan konten yang menyerupai karya berhak cipta yang dikenal?
- ➢ Apakah pengguna diperingatkan terhadap penggunaan output yang tidak semestinya?
- ➢ Apakah organisasi siap untuk menanggapi penghapusan atau litigasi?
- ➢ Apakah organisasi memiliki kebijakan dan praktik yang dapat dipertahankan?

Menanamkan Perlindungan Kontraktual

Karena banyak sistem AI dibuat oleh satu entitas dan digunakan oleh entitas lain, perjanjian hukum antara entitas-entitas ini perlu mengatasi risiko IP melalui penyertaan klausul-klausul seperti:

- ❖ Alokasi tanggung jawab yang jelas antara vendor dan pelanggan Klausul ganti rugi untuk klaim IP pihak ketiga
- ❖ Pembatasan penggunaan (misalnya, tidak menggunakan input dan/atau output untuk media komersial)
- ❖ Pernyataan dan jaminan mengenai sumber data pelatihan
- ❖ Ketentuan-ketentuan ini semakin umum dalam kontrak AI-as-a-Service (AIaaS), terutama untuk alat generatif yang menghasilkan konten kreatif.

Sebagai contoh perlindungan kontraktual, model generatif Firefly milik Adobe secara eksplisit menolak penggunaan data pihak ketiga yang dilindungi hak cipta dan membatasi hak pengguna untuk mengkomersialkan output. Adobe memberikan ganti rugi terbatas, sebuah tren yang menandakan peningkatan peran kontrak dalam mengelola risiko IP.

Berinvestasi dalam Kemampuan Penjelasan dan Transparansi

Model yang dapat menjelaskan outputnya atau melacak asal usul datanya menawarkan kemampuan pembelaan yang lebih besar dalam sengketa hukum terkait IP. Solusi yang muncul meliputi metode AI yang dapat dijelaskan (XAI), alat audit data sintetis, dan penandaan atau

hashing model untuk melacak penggunaan data. Meskipun tidak sepenuhnya sempurna, alat-alat ini membantu organisasi menunjukkan ketelitian dan dapat mengurangi risiko pelanggaran "kotak hitam".

Menurut Pendapat Penulis

Sistem AI yang kurang dapat dijelaskan mungkin hanya memiliki penggunaan terbatas di lingkungan komersial di mana kekayaan intelektual berperan, baik kekayaan intelektual tersebut digunakan sebagai input maupun diciptakan sebagai output. Klaim "kami tidak tahu bagaimana AI menciptakan ini" kemungkinan besar tidak akan berhasil sebagai bagian dari pembelaan hukum. Akibatnya, preseden hukum dapat mengesampingkan sistem AI yang tidak transparan.

Konflik Hak Kekayaan Intelektual Lintas Batas dan Kompleksitas Yurisdiksi

Sistem AI sering beroperasi melintasi batas negara: dirancang di satu negara, dilatih di negara lain, diterapkan di negara lain, dan diakses secara global. Namun, hukum hak kekayaan intelektual tetap bersifat teritorial: hukum tersebut diberikan dan ditegakkan oleh yurisdiksi masing-masing, sehingga menciptakan tantangan kepatuhan dan litigasi yang signifikan bagi operasi AI multinasional.

Standar Hukum yang Berbeda di Berbagai Yurisdiksi

Hukum mengenai apa yang dianggap sebagai pelanggaran, siapa yang dapat memegang hak cipta, dan bagaimana penggunaan wajar atau pengecualian berlaku sangat berbeda. Tabel 5.1 memberikan contohnya.

Tabel 5.1
Perbedaan Hukum Kekayaan Intelektual di Beberapa Yurisdiksi Terpilih

Wilayah	Perbedaan Utama
Amerika Serikat	Doktrin penggunaan wajar (*fair use*) mengizinkan penggunaan tanpa izin yang terbatas. Tidak ada hak cipta yang diberikan pada karya yang dihasilkan sepenuhnya oleh AI tanpa campur tangan manusia.
Uni Eropa	Pengecualian hak cipta yang lebih terbatas. Pengakuan hak basis data. Hak moral lebih kuat. Pengecualian penambangan teks dan data ada, tetapi mungkin memerlukan persetujuan dari pemegang hak.
Cina	Pengakuan eksplisit atas hak cipta dalam konten yang dihasilkan AI (di bawah bimbingan manusia). Keterlibatan negara dalam kepatuhan model.
Kanada/Australia	Menangani permasalahan berdasarkan kasus per kasus. Beberapa pihak terbuka terhadap pengecualian yang lebih luas untuk keperluan pelatihan.
Jepang	Mengizinkan penambangan teks dan data secara luas untuk tujuan apa pun, termasuk penggunaan komersial, asalkan tidak melanggar privasi atau keamanan. Belum ada pengakuan eksplisit atas kepengarangan AI.

Perancang dan pengembang sistem AI harus memperhitungkan perbedaan ini tidak hanya dalam penerapan model tetapi juga dalam akuisisi dataset, akses model, dan distribusi output.

Sebagai tanggapan terhadap ancaman hukum di Uni Eropa, beberapa alat AI generatif telah menerapkan geofencing (untuk membatasi akses pengguna dari negara dan wilayah tertentu) dan mengecualikan dataset tertentu dari penerapan khusus wilayah. Praktik ini mencerminkan kebutuhan untuk menyesuaikan layanan AI agar sesuai dengan standar hukum setempat.

Penegakan Yurisdiksi dan Pemilihan Forum

Ketika sengketa muncul, seperti seorang seniman yang mengklaim karyanya digunakan dalam pelatihan, penggugat dapat meneliti yurisdiksi yang menguntungkan, yang mengarah pada pemilihan forum, di mana proses penemuan dan gugatan kelompok menawarkan keuntungan strategis.

Organisasi multinasional (yang secara wajar mencakup organisasi yang beroperasi di satu negara dengan pengguna di banyak negara) harus mempersiapkan diri untuk:

- Klaim IP multi-yurisdiksi.
- Konflik hukum: Misalnya, di mana pelanggaran dianggap telah terjadi?
- Penegakan ekstrateritorial: Risiko tanggung jawab bahkan tanpa kehadiran fisik jika sistem AI atau outputnya ditawarkan di pasar tertentu.
- Pengungkapan data yang dipaksakan selama penemuan atau peninjauan peraturan.

Tantangan hukum ini dapat meningkatkan hambatan ekonomi untuk memasuki pasar vendor AI baru, berdasarkan kebutuhan akan penasihat hukum dengan pengalaman kekayaan intelektual global.

Perjanjian Internasional tentang Kekayaan Intelektual

Perjanjian seperti Konvensi Berne tahun 1886 dan Perjanjian TRIPS tahun 1994 memberikan perlindungan kekayaan intelektual dasar di lebih dari 170 negara. Meskipun mereka menyelaraskan beberapa hak kekayaan intelektual, mereka tidak menyelesaikan pertanyaan inti terkait AI, seperti:

- Apakah pelatihan pada data yang dilindungi hak cipta merupakan "penggunaan wajar," "reproduksi," atau sesuatu yang lain?
- Bisakah karya yang dihasilkan mesin dilindungi hak cipta?
- Bagaimana organisasi manajemen kolektif (CMO) melisensikan hak pelatihan?

Kesenjangan yang terkait dengan masukan dan keluaran dari sistem AI membutuhkan instrumen global baru yang membahas masalah kekayaan intelektual khusus AI.

Masa Depan Regulasi AI

Seiring meningkatnya litigasi terkait kekayaan intelektual, regulator mulai mempertimbangkan apakah undang-undang yang ada sudah cukup atau apakah kerangka kerja baru diperlukan untuk mengatur AI (menurut penulis, hal ini sudah terlambat). Bagian ini mengeksplorasi sinyal awal pergerakan legislatif.

Usulan Legislasi AS

Beberapa anggota Kongres AS telah mengajukan proposal untuk:

- Mewajibkan pengembang AI untuk mendokumentasikan sumber data
- Melarang penggunaan tanpa izin (yang, dalam beberapa kasus, masih perlu didefinisikan) dari konten berhak cipta untuk pelatihan

- Membangun registri atau kerangka kerja perizinan untuk kumpulan data pelatihan
- Menghalangi negara bagian AS untuk memberlakukan regulasi AI mereka sendiri, yang menyebabkan kebingungan seperti yang dialami dengan undang-undang privasi banyak negara bagian

Meskipun belum ada undang-undang federal komprehensif yang disahkan, peningkatan pengawasan dari Kantor Hak Cipta, FTC, dan pengadilan dapat mendorong pembuatan kebijakan di masa mendatang.

Perintah eksekutif terkait AI yang dikeluarkan oleh Presiden AS telah berfokus pada inovasi dan kepemimpinan AS yang berkelanjutan dalam teknologi AI, tetapi sejauh ini belum ada yang membahas hak kekayaan intelektual.

Yang Perlu Diperhatikan: Kasus Kekayaan Intelektual Masa Depan yang Dapat Membentuk Kembali Hukum AI
• Disney & Universal v. Midjourney • Getty Images v. Stability AI (Amerika Serikat dan Inggris) • Andersen v. Stability AI (Amerika Serikat) • Litigasi GitHub Copilot (Amerika Serikat) • Pembahasan Parlemen Eropa tentang pengungkapan konten yang dihasilkan AI

Kasus-kasus ini dapat menetapkan preseden tentang apakah pelatihan pada data berhak cipta merupakan pelanggaran, siapa yang bertanggung jawab atas output, dan kewajiban pengungkapan apa yang dihadapi pengembang.

Inisiatif Uni Eropa

Meskipun Undang-Undang AI Uni Eropa (dibahas dalam Bab 6) tidak secara langsung menyelesaikan masalah IP, undang-undang ini memperkenalkan kewajiban transparansi, seperti mendokumentasikan sumber data pelatihan dan mengungkapkan apakah data pelatihan berhak cipta digunakan.

Secara terpisah, Undang-Undang Layanan Digital Uni Eropa (DSA) dan Arahan Hak Cipta Uni Eropa dapat memengaruhi platform AI yang mendistribusikan konten. Regulasi ini dapat mewajibkan mekanisme penghapusan untuk output yang melanggar, mengenakan tanggung jawab pada platform yang menampung alat generatif, dan mendorong kerangka kerja perizinan untuk tujuan pembelajaran mesin. Hal ini akan memberikan kejelasan yang dibutuhkan, dan mudah-mudahan tidak terlalu membatasi kreativitas dan inovasi sistem AI.

Standar Industri dan Regulasi Mandiri

Tanpa mandat hukum yang jelas, industri AI mulai mengembangkan norma-normanya sendiri. Inisiatif sukarela meliputi:

- ✓ Kartu model yang mengungkapkan sumber dan batasan data
- ✓ Label nutrisi data yang menguraikan asal usul konten dan asal-usul data
- ✓ Pasar lisensi AI (misalnya, model data terbuka LAION [Jaringan Terbuka Kecerdasan Buatan Skala Besar, sebuah organisasi nirlaba Jerman], portal lisensi AI Shutterstock)

Organisasi yang mengikuti praktik terbaik yang muncul, terutama tentang persetujuan, atribusi, dan transparansi, lebih mungkin untuk bertahan dari pengawasan hukum di masa depan. Para optimis percaya bahwa industri apa pun yang cukup dan efektif mengawasi dirinya sendiri dapat menghindari peraturan yang keras, setidaknya untuk sementara waktu.

Hukum Kekayaan Intelektual sebagai Tantangan Tata Kelola Utama

Hukum kekayaan intelektual bukan lagi masalah khusus bagi pengembang AI, tetapi sebaliknya, ini adalah tantangan tata kelola utama. Karena sistem AI bergantung pada kumpulan data yang besar, seringkali tidak terkurasi, untuk menghasilkan keluaran yang kuat, risiko pelanggaran, sengketa kepemilikan, dan litigasi lintas batas semakin meningkat. Untuk unggul dalam persaingan, organisasi yang membangun atau menerapkan AI harus:

- Mendokumentasikan dan memeriksa sumber data pelatihan dengan cermat
- Memahami peraturan dan kewajiban IP global
- Mencantumkan klausul perlindungan risiko IP dalam perjanjian hukum
- Memantau litigasi dan perkembangan regulasi
- Mencantumkan aktivitas ini dalam kebijakan dan praktik yang dipersyaratkan oleh tata kelola AI

Dengan melakukan hal tersebut, tidak hanya membatasi risiko hukum tetapi juga menumbuhkan kepercayaan dengan pengguna, pencipta, dan regulator.

AI Generatif dan Posisi Kantor Hak Cipta

Kantor Hak Cipta AS telah mengeluarkan beberapa pernyataan yang mengklarifikasi bahwa:

- ❖ Karya yang dibuat semata-mata oleh penulis non-manusia tidak dilindungi oleh hukum hak cipta
- ❖ Pemohon harus mengungkapkan keterlibatan AI dalam pendaftaran hak cipta
- ❖ Kelalaian pengungkapan tersebut dapat membatalkan pendaftaran

Pembaruan kebijakan tahun 2023 meresmikan aturan ini, mendesak para pencipta untuk membedakan kontribusi manusia dan AI dalam karya dengan kepengarangan campuran, sebuah sinyal bahwa praktik pendaftaran IP akan menjadi lebih ketat di era AI.

5.2 HUKUM ANTI-DISKRIMINASI DAN AI

Seiring dengan semakin meluasnya penggunaan AI, organisasi-organisasi menggunakan sistem AI dalam pengambilan keputusan yang berisiko tinggi: siapa yang mendapatkan wawancara kerja, siapa yang diberhentikan dalam PHK perusahaan, berapa suku bunga yang diterima peminjam, apakah polis asuransi disetujui, apakah asuransi kesehatan akan menanggung biaya pengobatan, atau daftar perumahan mana yang memenuhi syarat bagi pelamar. Di setiap domain ini, hukum hak sipil dan anti-diskriminasi yang telah lama berlaku memberlakukan kewajiban hukum untuk memastikan keadilan dan perlakuan yang sama. Ketika AI terlibat, risiko melanggengkan atau memperburuk bias menjadi masalah kepatuhan hukum dan etika. Bagian ini mengeksplorasi bagaimana hukum anti-diskriminasi

berlaku untuk sistem AI, khususnya dalam bidang ketenagakerjaan, kredit, perumahan, pinjaman, dan asuransi, serta menawarkan panduan tentang mitigasi risiko dan tren regulasi.

Dasar Hukum Anti-Diskriminasi dalam Sistem Algoritma

Sebelum mengeksplorasi sektor industri tertentu, penting untuk memahami prinsip-prinsip hukum utama yang mendasari kewajiban anti-diskriminasi dalam sistem otomatis. Prinsip-prinsip ini umumnya berlaku di berbagai bidang dan menjadi tulang punggung tindakan penegakan hukum dan litigasi.

Kelompok yang Dilindungi dan Perilaku yang Dilarang

Hukum hak-hak sipil AS melarang diskriminasi berdasarkan "karakteristik yang dilindungi," yang umumnya mencakup kelompok-kelompok yang dilindungi berikut ini:

- Ras
- Warna kulit
- Asal negara
- Jenis kelamin (termasuk identitas gender dan orientasi seksual)
- Kehamilan
- Pendapatan
- Status keluarga
- Agama
- Disabilitas
- Status veteran
- Sifat genetik
- Usia

Perilaku yang dilarang biasanya terbagi dalam dua kategori:

- **Perlakuan yang tidak adil:** Diskriminasi yang disengaja berdasarkan kelompok yang dilindungi
- **Dampak yang tidak adil:** Kebijakan netral yang secara tidak proporsional merugikan kelompok yang dilindungi

Dasar-Dasar Dampak Diskriminatif

Suatu kebijakan atau model yang tampak netral tetapi secara tidak proporsional memengaruhi kelompok yang dilindungi dapat melanggar undang-undang anti-diskriminasi, bahkan jika tidak ada niat untuk melakukan diskriminasi yang ditunjukkan. Penggugat harus menunjukkan:

- Dampak buruk yang signifikan terhadap kelompok yang dilindungi
- Kurangnya kebutuhan bisnis atau alternatif yang kurang diskriminatif

Teori dampak diskriminatif merupakan inti dari banyak kasus bias algoritmik.

Bias Algoritma sebagai Risiko Hukum

Bias dalam sistem AI, baik disengaja maupun tidak, dapat muncul dari:

- Data pelatihan yang bias (misalnya, keputusan historis yang mencerminkan diskriminasi sistemik)
- Variabel proksi (misalnya, kode pos yang berkorelasi dengan ras atau agama)
- Model buram yang menutupi bias
- Lingkaran umpan balik yang memperkuat ketidaksetaraan
- Bias pengambilan sampel (kumpulan data yang kurang mewakili atau terlalu mewakili kelompok atau konteks tertentu)

Ketika masalah ini muncul dalam perekrutan, pemberian pinjaman, perumahan, atau asuransi, hal itu dapat melanggar undang-undang anti-diskriminasi dan membuat organisasi rentan terhadap tuntutan hukum atau sanksi peraturan. Ketika sistem AI memfasilitasi pengambilan keputusan, organisasi mungkin mendapati diri mereka membela sistem tersebut serta desain dan pelatihannya.

Pekerjaan: AI dalam Perekrutan, Promosi, dan Pemutusan Hubungan Kerja

Sistem AI semakin banyak digunakan dalam keputusan ketenagakerjaan, mulai dari penyaringan resume dan wawancara video hingga analitik tenaga kerja dan manajemen kinerja. Meskipun demikian, alat-alat ini harus mematuhi hukum ketenagakerjaan dan hak-hak sipil, khususnya Pasal VII Undang-Undang Hak Sipil, Undang-Undang Penyandang Disabilitas Amerika (ADA), dan Undang-Undang Diskriminasi Usia dalam Pekerjaan (ADEA). Hanya karena sistem AI memfasilitasi proses ini, apakah sistem AI tersebut transparan atau tidak, tidak membebaskan organisasi mana pun dari tanggung jawab ketika terjadi diskriminasi.

Persyaratan Hukum untuk Algoritma Perekrutan

Berdasarkan Pasal VII Undang-Undang Hak Sipil, adalah melanggar hukum bagi pemberi kerja untuk menggunakan prosedur seleksi yang menyebabkan dampak yang berbeda pada kelompok yang dilindungi kecuali:

- ✓ Prosedur tersebut terkait dengan pekerjaan dan konsisten dengan kebutuhan bisnis
- ✓ Tidak ada alternatif yang kurang diskriminatif yang tersedia

Aturan ini berlaku untuk keputusan yang digerakkan oleh manusia maupun mesin.

Pada tahun 2018, Amazon mengembangkan dan kemudian meninggalkan alat penyaringan resume internal yang menghukum kandidat dari perguruan tinggi wanita atau yang menggunakan istilah seperti "klub catur wanita."

Model tersebut telah dilatih pada resume dari kumpulan pelamar yang didominasi laki-laki, sehingga menimbulkan bias gender melalui data dan praktik historis. Meskipun tidak pernah diterapkan dalam skala besar, alat ini berfungsi sebagai pelajaran berharga tentang data pelatihan yang tidak diperiksa dan diskriminasi terselubung.

Panduan EEOC tentang AI dalam Ketenagakerjaan

Pada Mei 2023, Komisi Kesempatan Kerja yang Setara (EEOC) mengeluarkan panduan yang mengklarifikasi bahwa pemberi kerja bertanggung jawab atas keputusan berbasis AI, bahkan jika vendor pihak ketiga mengembangkan alat AI tersebut. Panduan ini konsisten dengan prinsip umum manajemen risiko pihak ketiga. Jika produk atau layanan vendor atau penyedia layanan mana pun mengakibatkan suatu peristiwa atau insiden, organisasi yang memilih vendor tersebut pada akhirnya bertanggung jawab atas insiden tersebut.

Poin-Poin Penting dari EEOC
➢ Perusahaan harus mengevaluasi alat-alat yang digunakan untuk mendeteksi dampak yang tidak proporsional. ➢ Penafian dari vendor tidak melindungi tanggung jawab hukum. ➢ Akomodasi yang wajar harus diberikan berdasarkan ADA, bahkan untuk penilaian algoritmik. ➢ Perusahaan harus mengaudit dan mendokumentasikan proses pengambilan keputusan algoritmik.
Panduan ini mencerminkan konsensus yang berkembang bahwa "pengalihan algoritmik" terhadap bias bukanlah pembelaan.

Hukum Lokal dan Peraturan Perekrutan yang Adil

Di luar hukum federal, kota dan negara bagian memberlakukan peraturan ketenagakerjaan terkait AI mereka sendiri. Misalnya, Hukum Kota New York 144 mewajibkan audit bias terhadap alat pengambilan keputusan ketenagakerjaan otomatis, publikasi tahunan hasil audit, dan pengungkapan kepada kandidat tentang penggunaan algoritma. Hukum Kota New York adalah yang pertama dari jenisnya di Amerika Serikat, tetapi hukum serupa sedang diusulkan di California, Illinois, dan yurisdiksi lainnya. Pada saat Anda membaca ini, kemungkinan akan ada beberapa lagi, terutama karena Kongres tidak mengesahkan hukum "tidak ada hukum negara bagian tentang AI".

Kredit dan Pinjaman: Menavigasi FCRA dan ECOA di Era AI

Keputusan kredit adalah wilayah utama untuk aplikasi AI, mulai dari penjaminan pinjaman dan deteksi penipuan hingga penilaian kredit dan penetapan harga dinamis. Namun, aktivitas ini diatur secara ketat di Amerika Serikat oleh Undang-Undang Kesempatan Kredit yang Setara (ECOA) dan Undang-Undang Pelaporan Kredit yang Adil (FCRA), yang keduanya berlaku terlepas dari apakah manusia atau mesin yang membuat keputusan.

Organisasi yang mempertimbangkan penggunaan AI untuk meningkatkan pengambilan keputusan dalam bisnis kredit dapat dikatakan optimis tentang peningkatan portofolio pinjaman atau pendapatan bersih mereka, dan bahwa peningkatan tersebut akan sepadan dengan berbagai jenis risiko yang ditimbulkan oleh proyek TI apa pun (apalagi AI).

Undang-Undang Kesempatan Kredit yang Setara (ECOA)

Undang-Undang Kesempatan Kredit yang Setara (ECOA) tahun 1974 melarang diskriminasi dalam transaksi kredit berdasarkan karakteristik yang dilindungi, termasuk ras, jenis kelamin, status perkawinan, agama, asal negara, usia, atau penerimaan bantuan publik. Undang-undang ini berlaku untuk kartu kredit, hipotek, pinjaman mobil, pinjaman usaha kecil, dan jalur kredit pribadi.

Ketika sistem AI digunakan untuk membuat atau membantu pengambilan keputusan ini, kreditor harus memastikan:

❖ Tidak ada dampak yang tidak proporsional atau penggunaan proksi terlarang

- ❖ Ada alasan yang jelas untuk tindakan yang merugikan (kemampuan menjelaskan AI)
- ❖ Akses yang adil di seluruh kelompok yang dilindungi

Pada tahun 2022, Biro Perlindungan Keuangan Konsumen (CFPB) memperingatkan pemberi pinjaman bahwa algoritma AI "kotak hitam" tidak membebaskan mereka dari kewajiban ECOA. Bahkan jika modelnya buram, pemberi pinjaman harus dapat memberikan alasan spesifik untuk penolakan kredit. Aturan ini adalah alasan lain untuk memastikan bahwa setiap model AI yang dikembangkan memiliki kemampuan menjelaskan yang solid, karena undang-undang kemungkinan akan mensyaratkannya.

Undang-Undang Pelaporan Kredit yang Adil (FCRA)

Undang-Undang Pelaporan Kredit yang Adil (FCRA) AS tahun 1970 mengatur penggunaan skor kredit dan data pihak ketiga dalam pengambilan keputusan kredit. Ketika sistem AI menyerap atau menghasilkan data tersebut, mereka dapat memenuhi syarat sebagai lembaga pelaporan konsumen (CRA—juga dikenal sebagai biro kredit) berdasarkan FCRA.

FCRA mensyaratkan:

- Akurasi dan integritas data konsumen
- Pengungkapan kepada konsumen tentang keputusan dan sumber data
- Mekanisme untuk membantah kesalahan

FCRA juga berlaku untuk pemeriksaan latar belakang pekerjaan. Di banyak yurisdiksi, laporan kredit dapat digunakan sebagai bagian dari investigasi latar belakang yang digunakan banyak organisasi sebagai bagian dari proses penyaringan mereka.

FCRA vs. ECOA

Aspek	**FCRA**	**ECOA**
Fokus	Data dan laporan konsumen	Diskriminasi dalam pengambilan keputusan pemberian pinjaman
Pemicu	Penggunaan data konsumen pihak ketiga	Status kelas dilindungi
Kunci Persyaratan	Laporan yang akurat dan dapat diperdebatkan	Akses yang setara dan kriteria yang dapat dibenarkan
AI Implikasi	Sumber data yang bertanggung jawab	Logika deteksi dan justifikasi bias

Secara bersama-sama, hukum-hukum ini mensyaratkan input dan output yang adil dalam model AI yang berkaitan dengan kredit.

Bias dalam Model Penilaian Kredit

Bahkan ketika model AI tidak secara langsung menyertakan karakteristik yang dilindungi, mereka mungkin bergantung pada proksi yang berkorelasi (misalnya, kode pos, tingkat pendidikan, aktivitas jejaring sosial). Proksi ini dapat menyebabkan dampak yang tidak diinginkan dan berbeda.

Pada tahun 2019, beberapa pelanggan, termasuk salah satu pendiri Apple, Steve Wozniak, melaporkan menerima batas kredit yang jauh lebih rendah untuk kartu kredit Apple daripada pasangan mereka meskipun memiliki profil keuangan yang serupa. Insiden tersebut mendorong penyelidikan oleh Departemen Layanan Keuangan New York (NYDFS) tentang

apakah algoritma Goldman Sachs melanggar undang-undang pinjaman yang adil. Meskipun tidak ditemukan niat diskriminatif, kasus tersebut menggarisbawahi risiko ketidakjelasan algoritma dan kurangnya penjelasan dalam keputusan kredit.

Perumahan: AI dan Undang-Undang Perumahan yang Adil

AI sekarang digunakan oleh pemilik rumah, pengelola properti, dan platform perumahan untuk menyaring penyewa, menentukan harga sewa, merekomendasikan daftar properti, dan menganalisis daya tarik lingkungan. Sistem ini harus mematuhi Undang-Undang Perumahan yang Adil (FHA), yang melarang diskriminasi dalam penjualan, penyewaan, dan pembiayaan perumahan.

Karakteristik yang Dilindungi Berdasarkan FHA

Undang-Undang Perumahan Adil (*Fair Housing Act*/FHA) tahun 1968 merupakan bagian dari Undang-Undang Hak Sipil tahun 1968, dan melindungi dari diskriminasi berdasarkan beberapa karakteristik, termasuk:

- Ras
- Warna kulit
- Agama
- Jenis kelamin
- Asal negara
- Disabilitas
- Status keluarga (misalnya, status perkawinan, dan memiliki anak)

Perlindungan ini meluas ke platform penyaringan penyewa berbasis AI, sistem manajemen properti, dan algoritma iklan sewa dan hipotek.

Iklan dan Penargetan Algoritma

Platform digital sering menggunakan AI untuk mengoptimalkan iklan perumahan berdasarkan perilaku pengguna atau minat yang diprediksi (pemasaran yang ditargetkan bukanlah hal baru—teknologi digital hanya membuatnya lebih baik). Jika hal ini mengakibatkan paparan yang berbeda (di mana sistem AI menyebabkan pengguna dan kelompok menerima informasi yang berbeda, bahkan ketika sistem memperlakukan semua pengguna secara sama dalam hal logika pengambilan keputusan) berdasarkan kelas yang dilindungi, hal itu dapat melanggar FHA, bahkan jika tidak disengaja.

Pada tahun 2019, Departemen Perumahan dan Pembangunan Perkotaan AS (HUD) menuduh Facebook mengizinkan pengiklan perumahan untuk mengecualikan pengguna berdasarkan ras, agama, jenis kelamin, dan kode pos. Facebook dan HUD kemudian mencapai kesepakatan, dan Facebook menerapkan aturan baru untuk penargetan iklan di bidang perumahan, kredit, dan pekerjaan.

Apa yang Dianggap sebagai Pelanggaran?

Berdasarkan Undang-Undang Perumahan Adil AS, pelanggaran meliputi:

- Penargetan diskriminatif (misalnya, mengecualikan pengguna dari etnis tertentu)
- Dampak diskriminatif (misalnya, algoritma membatasi akses ke kelompok tertentu)
- Kegagalan untuk memantau alat pihak ketiga

Semua pelanggaran ini dapat memicu tanggung jawab dan sanksi. Penyedia perumahan dan platform harus mengaudit algoritma untuk memastikan keadilan di semua kelompok yang dilindungi.

Penyaringan Penyewa dan Bias Berbasis Kredit

Alat penyaringan penyewa berbasis AI dapat menggunakan skor kredit, riwayat kriminal, atau catatan pengusiran, yang semuanya dapat mencerminkan bias sistemik. Jika kriteria ini secara tidak proporsional mengecualikan kelompok yang dilindungi, pemilik rumah dapat menghadapi keluhan perumahan yang adil.

Beberapa tuntutan hukum telah menantang perusahaan penyaringan penyewa karena memasukkan data pengusiran yang tidak akurat atau usang yang secara tidak proporsional memengaruhi komunitas kulit berwarna. Pengadilan telah memutuskan bahwa metode penyaringan, termasuk yang diaktifkan oleh alat otomatis, harus memenuhi standar akurasi, relevansi, dan non-diskriminasi berdasarkan FHA dan FCRA.

Asuransi: Penilaian Risiko dalam Lanskap yang Teregulasi

Industri asuransi telah lama bergantung pada analitik data untuk menilai risiko dan menetapkan harga. Dengan munculnya AI, perusahaan asuransi sekarang menggunakan model yang menganalisis data perilaku, aktivitas media sosial, sensor IoT, dan bahkan gambar wajah—semuanya untuk meningkatkan akurasi data aktuaria mereka. Meskipun kemampuan ini menjanjikan penetapan harga yang dipersonalisasi, hal itu dilakukan dengan risiko memperkenalkan diskriminasi yang melanggar hukum.

Tidak seperti pemeriksaan kredit dan pekerjaan, tidak ada satu pun undang-undang anti-diskriminasi federal untuk asuransi. Regulasi asuransi terutama terjadi di tingkat negara bagian AS, dengan berbagai perlindungan dan mekanisme pengawasan.

Diskriminasi dalam Penjaminan dan Penetapan Harga Asuransi

Diskriminasi asuransi dapat terjadi melalui penjaminan (melalui penolakan atau pembatasan cakupan) dan penetapan harga (dengan mengenakan premi yang berbeda kepada individu yang berada dalam situasi serupa). Undang-undang negara bagian biasanya melarang diskriminasi yang tidak adil berdasarkan ras, agama, jenis kelamin, disabilitas, asal kebangsaan atau etnis, status veteran, dan informasi genetik. Beberapa negara bagian telah memperluas perlindungan untuk mencakup orientasi seksual, identitas gender, dan pekerjaan. Penulis ini tidak akan terkejut jika melihat perlindungan tambahan berdasarkan konten media sosial.

Beberapa negara bagian (California, Hawaii, dan Massachusetts) melarang atau membatasi penggunaan skor kredit dalam penjaminan asuransi mobil karena kekhawatiran akan bias rasial. Model AI yang menggunakan proksi tersebut dapat melanggar larangan ini, bahkan jika tidak disengaja.

Diskriminasi Proksi dan Pemilihan Fitur

Risiko umum dalam AI asuransi adalah penggunaan variabel proksi: faktor-faktor yang berkorelasi dengan karakteristik yang dilindungi tetapi tidak secara langsung dilarang.

Variabel Proksi dalam Asuransi

Variabel	**Proksi Potensial Untuk**
Kode Pos	Ras, status sosial ekonomi, agama
Tingkat pendidikan	Pendapatan atau kewarganegaraan
Pekerjaan	Jenis kelamin atau status kelas
Sejarah kredit	Ras (melalui ketidaksetaraan sistemik)

Perusahaan asuransi harus membenarkan fitur model berdasarkan relevansi aktuaria, bukan hanya korelasi semata. Negara bagian seperti Colorado telah mewajibkan perusahaan asuransi untuk menunjukkan bahwa model AI mereka tidak menghasilkan hasil yang berbeda-beda.

Perkembangan AI di Tingkat Negara Bagian dalam Asuransi

Sebagai regulator utama untuk asuransi, negara-negara bagian AS memperkenalkan undang-undang atau panduan tentang keadilan AI dalam asuransi. Misalnya, SB21-169 Colorado (melindungi konsumen dari diskriminasi yang tidak adil dalam praktik asuransi) menjadi undang-undang negara bagian pertama yang mewajibkan perusahaan asuransi untuk:

- Mengevaluasi sistem AI dan big data untuk diskriminasi yang tidak adil
- Menyerahkan dokumentasi pengujian bias
- Menerapkan kebijakan tata kelola untuk mengurangi dampak diskriminatif

Model ini dapat menginspirasi undang-undang serupa di yurisdiksi lain, meningkatkan standar akuntabilitas algoritma. Pada saat buku ini dicetak, diharapkan lebih banyak negara bagian AS akan memberlakukan peraturan AI.

Strategi Mitigasi Risiko dan Kepatuhan

Di setiap sektor bisnis dengan peraturan diskriminasi, organisasi harus membangun mitigasi risiko hukum ke dalam kerangka kerja tata kelola AI mereka (tidak ada yang baru di sini, karena setiap kerangka kerja tata kelola harus mencakup manajemen risiko). Bagian ini menguraikan langkah-langkah untuk mematuhi undang-undang anti-diskriminasi yang ada dan yang diantisipasi sambil memungkinkan inovasi yang didukung AI.

Melakukan Audit Bias dan Penilaian Dampak

Audit bias sangat penting untuk mengidentifikasi dampak yang tidak merata sebelum penerapan. Elemen kunci dari audit bias meliputi:

- ✓ Evaluasi statistik dari dataset pelatihan dan pengujian
- ✓ Penilaian kinerja model di seluruh subkelompok demografis
- ✓ Tinjauan relevansi fitur dan risiko proksi
- ✓ Dokumentasi metodologi audit, hasil audit, keputusan mitigasi, dan hasilnya
- ✓ Pemantauan pasca-penerapan yang berkelanjutan untuk mendeteksi pergeseran bias

Daftar Periksa Bias Model AI

Beberapa hal yang perlu diperhatikan dalam audit model AI meliputi:

- ➢ Apakah metrik dampak yang berbeda telah diuji
- ➢ Apakah proksi kelas yang dilindungi telah dihapus atau dibenarkan
- ➢ Apakah ada kebutuhan bisnis untuk fitur yang berdampak

- Apakah keluaran model dapat dijelaskan dan ditinjau

Sekali lagi, kita melihat alasan yang kuat bagi organisasi untuk hanya menggunakan AI yang dapat dijelaskan dan transparan.

Pastikan Keterjelasan dan Transparansi

Kepatuhan hukum seringkali bergantung pada apakah suatu organisasi dapat menjelaskan mengapa sistem AI membuat keputusan, terutama dalam tindakan yang merugikan. Praktik penting meliputi:

- Gunakan model yang dapat diinterpretasikan atau alat penjelasan pasca-keputusan
- Berikan alasan yang dapat dipertahankan dan ditindaklanjuti untuk penolakan atau hasil yang merugikan
- Catat keputusan dan perilaku model dari waktu ke waktu
- Tinjau perilaku model secara berkala dari waktu ke waktu

Tetapkan Pengawasan dan Akuntabilitas Manusia

Keputusan AI, khususnya yang memengaruhi hak subjek individu, tidak boleh terjadi dalam kekosongan tata kelola. Kerangka kerja tata kelola AI yang kuat harus mencakup:

- Tinjauan dengan melibatkan manusia untuk keputusan berisiko tinggi dan bernilai tinggi
- Kepemilikan yang jelas atas kepatuhan AI dalam tim hukum atau kepatuhan
- Kurasi proaktif terhadap peraturan terkait AI yang muncul, disertai dengan tinjauan kepatuhan
- Pelatihan rutin tentang hukum hak sipil dan etika AI
- Koordinasi antara staf teknis, hukum, dan operasional
- Laporan tata kelola berkala kepada eksekutif senior

Sertakan Klausul Keadilan dalam Kontrak Vendor

Banyak organisasi bergantung pada satu atau lebih pihak ketiga untuk menyusun kemampuan AI mereka yang membuat atau membantu dalam pengambilan keputusan bisnis. Perlindungan kontraktual sangat penting untuk mendefinisikan, membagi, dan menetapkan tanggung jawab serta memastikan kepatuhan.

Selain praktik manajemen risiko pihak ketiga yang ada terkait perjanjian hukum, ketentuan kontrak harus mewajibkan vendor untuk:

- Melakukan dan membagikan hasil audit bias
- Mengungkapkan sumber data dan dokumentasi model
- Menyetujui untuk mengganti kerugian atas pelanggaran peraturan
- Membantu dengan penjelasan dan pemberitahuan tindakan yang merugikan

Banyak departemen hukum organisasi menggunakan templat perjanjian hukum, yang harus diperbarui secara berkala agar tetap sesuai dengan hal-hal kontraktual ini dan lainnya.

Keadilan sebagai Mandat Hukum, Bukan Hanya Tujuan Etis

Keadilan dalam sistem AI bukan hanya ide yang bagus: itu adalah hukum. Sistem AI yang beroperasi di bidang ketenagakerjaan, keuangan, perumahan, dan asuransi tidak berada di luar hukum. Undang-undang anti-diskriminasi yang telah lama berlaku, dari Title VII hingga Fair Housing Act, berlaku terlepas dari apakah manusia atau mesin yang membuat keputusan.

Organisasi yang mempertimbangkan untuk menerapkan AI dalam konteks sensitif ini harus:

- Mengevaluasi secara realistis keuntungan ekonomi atau kompetitif yang akan dibawa oleh sistem AI
- Memahami dan mematuhi hukum yang berlaku
- Mengantisipasi peraturan baru
- Mengaudit model untuk dampak yang berbeda dan bias proksi
- Memberikan transparansi dan penjelasan kepada individu yang terkena dampak
- Mendokumentasikan prosedur tata kelola untuk menunjukkan manajemen proaktif dan itikad baik

Keadilan dalam AI bukan hanya keharusan moral atau etis; itu adalah kewajiban hukum, yang ditegakkan oleh regulator dan penggugat.

Tanda Bahaya yang Memicu Pengawasan Hukum

Regulator, pengadilan, dan kelompok advokasi dapat fokus pada sistem AI yang diamati:

- Mengandalkan model yang tidak transparan dan kurang dapat dijelaskan
- Menggunakan riwayat kredit, kode pos, atau pendidikan tanpa justifikasi yang wajar
- Tidak mampu memberikan alasan spesifik untuk keputusan yang merugikan
- Diterapkan tanpa pengujian dampak yang berbeda
- Menolak permintaan akomodasi yang wajar

Menghindari jebakan ini dimulai dengan menanamkan kepatuhan hukum ke dalam desain, pengembangan, dan penerapan AI sejak awal. Tinjauan tujuan bisnis dan desain sebelum pengembangan membantu organisasi menghindari masalah ini dan masalah lainnya.

5.3 BAGAIMANA HUKUM PERLINDUNGAN KONSUMEN BERLAKU UNTUK AI

Seiring semakin terintegrasinya AI dalam produk dan layanan konsumen, AI semakin memengaruhi cara kita berbelanja, belajar, mendapatkan layanan kesehatan, dan mengelola keuangan kita. Ketika sistem ini menyesatkan, menipu, memanipulasi, atau merugikan pengguna, sistem tersebut dapat melanggar larangan hukum terhadap tindakan atau praktik yang tidak adil atau menipu (UDAP), yang dikenal juga sebagai iklan palsu atau menyesatkan. Bagian ini membahas bagaimana hukum perlindungan konsumen AS berlaku untuk sistem AI dan memberikan panduan praktis untuk tata kelola, transparansi, dan mitigasi risiko.

Dasar Hukum: Tindakan atau Praktik yang Tidak Adil dan Menipu

Hukum perlindungan konsumen di Amerika Serikat berasal dari undang-undang federal dan negara bagian yang melarang perilaku yang tidak adil, menipu, dan curang dalam perdagangan. Hukum-hukum ini sengaja dibuat luas dan fleksibel, menjadikannya alat yang ampuh untuk mengatur perilaku AI.

Undang-Undang FTC dan Wewenang Bagian 5

Komisi Perdagangan Federal AS (FTC) adalah penegak hukum federal utama untuk perlindungan konsumen. Berdasarkan Bagian 5 Undang-Undang FTC, lembaga ini dapat mengambil tindakan hukum terhadap:

- ✓ **Tindakan atau praktik yang tidak adil:** Tindakan atau praktik yang menyebabkan atau kemungkinan besar menyebabkan kerugian besar bagi konsumen, tidak dapat dihindari secara wajar, dan tidak diimbangi oleh manfaat
- ✓ **Tindakan atau praktik yang menipu:** Tindakan atau praktik yang menyesatkan atau kemungkinan besar menyesatkan konsumen yang wajar, berdasarkan representasi atau kelalaian material

Standar ini berlaku terlepas dari apakah kerugian tersebut timbul dari taktik penjualan manusia, situs web, atau model AI.

UDAP vs. Penipuan
Tindakan dan praktik yang tidak adil dan menipu (UDAP) berbeda dari penipuan sebagai berikut:

Tindakan Tidak Adil atau Menipu	Tipuan
Tanggung jawab perdata	Seringkali kriminal
Tidak diperlukan niat	Membutuhkan bukti niat
Penerapan yang luas	Perilaku yang sempit dan spesifik
Ditegakkan oleh FTC, negara bagian	Dituntut oleh Departemen Kehakiman, Jaksa Agung, atau penggugat

Undang-Undang Perlindungan Konsumen Tingkat Negara Bagian

Setiap negara bagian AS memberlakukan versi undang-undang UDAP mereka sendiri, yang kadang-kadang disebut sebagai "Undang-Undang mini-FTC." Undang-undang ini sering memberikan perlindungan tambahan, definisi kerugian yang lebih luas, dan hak gugatan perdata. Contohnya meliputi:

- ➢ Undang-Undang Persaingan Tidak Adil California (UCL) banyak mengambil dari Undang-Undang FTC dan melarang "setiap tindakan atau praktik bisnis yang melanggar hukum, tidak adil, atau curang."
- ➢ Undang-Undang Perlindungan Konsumen Massachusetts 93A memungkinkan konsumen untuk mendapatkan ganti rugi tiga kali lipat untuk perilaku menipu tertentu.
- ➢ Undang-Undang Bisnis Umum New York (GBL) §349 berfokus pada perilaku menipu.

Undang-undang ini menciptakan risiko yang beragam bagi perusahaan yang menerapkan alat AI di berbagai yurisdiksi. Karena tidak ada larangan federal AS terhadap undang-undang AI tingkat negara bagian, keragaman ini diperkirakan akan menjadi lebih kompleks dalam beberapa tahun ke depan. Jika sejarah menjadi acuan (memikirkan undang-undang privasi federal dan negara bagian), kecil kemungkinan undang-undang AI federal yang menyeluruh akan menggantikan undang-undang negara bagian AS.

Area Fokus Baru untuk Penegakan AI

Otoritas perlindungan konsumen menyadari tren organisasi yang mengintegrasikan kemampuan AI ke dalam produk dan layanan mereka. Regulator semakin prihatin dengan:

- ❖ **Penggunaan AI yang tidak diungkapkan:** Konsumen yang berinteraksi dengan sistem AI mungkin tidak menyadari bahwa mereka tidak berurusan dengan manusia (sejauh ini, peraturan belum secara universal mewajibkan pengungkapan ini, namun mungkin etis untuk melakukannya).

- **Desain manipulatif:** AI yang mendorong, menipu, atau memaksa konsumen untuk berperilaku yang seharusnya tidak mereka pilih.
- **Konten yang tidak akurat atau berbahaya:** Chatbot atau mesin rekomendasi yang menghasilkan informasi palsu atau tidak aman.
- **Pengambilan keputusan yang tidak transparan:** Sistem AI yang memengaruhi kelayakan atau harga tetapi tidak menawarkan penjelasan atau jalan keluar.
- **Privasi dan keamanan data:** Sistem AI mengumpulkan, menyimpulkan, atau membagikan lebih banyak data pribadi daripada yang diperlukan atau dengan cara yang tidak disetujui konsumen.

Kekhawatiran ini mengarah pada tindakan penegakan hukum baru, dokumen panduan, dan usulan kebijakan—semuanya bertujuan untuk meminta pertanggungjawaban pengembang dan penyebar AI.

Penipuan atau Praktik Pelayanan Unggul

Hampir semua orang familiar dengan "umpan" media sosial mereka, serta peluang menonton dan membeli dari situs streaming (misalnya, Netflix, Prime Video) dan situs e-commerce (misalnya, Amazon, Shopify, Squarespace). Sebagian besar, jika tidak semua, menggunakan semacam algoritma ML atau AI untuk memilih apa yang kita lihat. Apakah praktik seperti itu dianggap manipulatif (karena kita mungkin "ditipu" untuk membeli lebih banyak) atau hanya praktik pelayanan unggul dengan menunjukkan kepada kita apa yang lebih mungkin kita inginkan?

Penegakan dan Kebijakan FTC di Era AI

Seperti halnya untuk privasi konsumen, FTC telah memposisikan dirinya sebagai pengawas AI utama, memanfaatkan wewenang perlindungan konsumen yang ada untuk mengawasi kerugian yang disebabkan oleh AI. Meskipun belum menerbitkan peraturan AI formal, lembaga tersebut telah mengeluarkan panduan dan melakukan tindakan penegakan hukum berdasarkan Bagian 5.

Panduan FTC tentang Keadilan dan Transparansi AI

Pada tahun 2021 dan 2022, FTC merilis panduan melalui postingan blog dan pernyataan kebijakan, memperingatkan perusahaan bahwa:

- Sistem AI yang menghasilkan hasil yang bias, tidak adil, atau menipu dapat melanggar hukum.
- Klaim tentang akurasi AI dan kemampuan AI harus dapat dibuktikan dan didukung bukti.
- Perusahaan bertanggung jawab atas perilaku AI, bahkan ketika menggunakan alat AI pihak ketiga (ini kembali ke praktik manajemen risiko pihak ketiga yang telah mapan).
- Kebersihan dan asal usul data sangat penting untuk pengembangan model yang sah.

Peringatan ini memperkuat pandangan FTC bahwa AI tidak dikecualikan dari pertanggungjawaban, hanya karena kompleks atau sulit dijelaskan.

Lima Pertanyaan Tata Kelola AI dari FTC

Menurut FTC, perusahaan harus menentukan apakah:

- Dataset AI mereka bias atau tidak lengkap
- Keputusan dan hasil yang didukung AI mereka mendiskriminasi kelompok yang dilindungi
- Konsumen dapat memilih untuk tidak ikut serta dalam keputusan otomatis
- Mereka transparan tentang kapan dan bagaimana AI digunakan
- Mereka dapat menjelaskan dan membenarkan keputusan dan hasil yang didukung AI mereka

Kegagalan untuk mengatasi masalah ini dapat menarik perhatian yang tidak diinginkan.

Tindakan Penegakan Hukum Utama FTC yang Melibatkan AI

Meskipun banyak kasus hukum diselesaikan tanpa pengakuan tanggung jawab, kasus-kasus tersebut menawarkan wawasan berharga tentang harapan regulasi.

Everalbum (2021)

Everalbum menawarkan aplikasi penyimpanan foto yang menggunakan pengenalan wajah tanpa memberi tahu penggunanya secara jelas. FTC menuduh perusahaan tersebut menyesatkan konsumen dengan mengaktifkan pengenalan wajah secara default tanpa pemberitahuan dan menyimpan data biometrik bahkan setelah pengguna menonaktifkan akun mereka. Everalbum diharuskan untuk menghapus model dan data pengenalan wajahnya dan secara tegas mengungkapkan praktik biometriknya di masa mendatang.

Aplikasi Penurunan Berat Badan "CuraLife" (2022)

Meskipun tidak sepenuhnya berbasis AI, kasus ini menargetkan klaim palsu tentang analisis kesehatan otomatis. CuraLife diduga memasarkan wawasan diabetes berbasis AI tanpa dukungan ilmiah yang memadai. Perusahaan harus mengungkapkan penggunaan AI dan membuktikan klaim apa pun tentang akurasi AI, manfaat kesehatan, atau personalisasi, sama seperti yang mereka lakukan untuk produk tradisional.

Amazon Alexa dan Data Anak-Anak (2019)

FTC dan DOJ menyelidiki dan mendenda produk Alexa Amazon karena melanggar Undang-Undang Perlindungan Privasi Online Anak-Anak (COPPA) dengan menyimpan rekaman suara anak-anak, gagal menghapus data yang diminta, dan tidak mengungkapkan secara memadai bagaimana Alexa menggunakan informasi tersebut. Meskipun bukan kasus AI secara ketat, contoh ini menggambarkan bagaimana asisten suara, pengenalan ucapan, dan penargetan yang dipersonalisasi dapat beririsan dengan perlindungan konsumen dan hukum privasi.

Penipuan dan Manipulasi yang Didukung AI

Tidak semua aplikasi AI menimbulkan kekhawatiran perlindungan konsumen, tetapi beberapa kasus penggunaan mungkin rentan terhadap manipulasi, informasi yang salah, atau penipuan. Ini termasuk model generatif, chatbot, mesin rekomendasi, dan alat bantu pengambilan keputusan. Bagian ini mengeksplorasi bagaimana hukum perlindungan

konsumen menanggapi sistem AI yang membentuk perilaku pengguna atau menyesatkan publik.

Pola Gelap dan Antarmuka Manipulatif

Seperti yang pertama kali dibahas di Bab 4, pola gelap mengacu pada desain antarmuka pengguna yang menipu, memaksa, atau membingungkan konsumen untuk membuat pilihan yang bertentangan dengan kepentingan terbaik mereka, seperti:

- Berlangganan layanan tanpa sengaja
- Berbagi lebih banyak data daripada yang mereka inginkan
- Mengklik opsi yang menyesatkan atau sudah dicentang sebelumnya
- Kesulitan untuk membatalkan atau berhenti berlangganan

AI memperburuk risiko ini dengan secara dinamis menyesuaikan konten berdasarkan prediksi perilaku waktu nyata, mengoptimalkan keterlibatan atau monetisasi daripada kesejahteraan konsumen, dan mempersonalisasi manipulasi dalam skala besar. Mengapa memanipulasi pelanggan satu per satu ketika AI dapat melakukannya dengan lebih efisien?!

Sebagai contoh, bayangkan layanan berbasis AI mempelajari bahwa pengguna larut malam lebih impulsif dan menyesuaikan antarmuka untuk menyoroti penawaran premium setelah pukul 10 malam. Jika dorongan ini menyebabkan konsumen berlangganan tanpa sengaja dan membuat pembatalan menjadi sulit, regulator dapat menganggap ini tidak adil atau menipu. Banyak dari kita telah mengalami layanan yang dapat dibeli hanya dengan beberapa klik. Namun, pembatalan membutuhkan navigasi yang panjang melalui menu penjawab otomatis yang membingungkan, bersama dengan waktu tunggu yang lama. Banyak yang akan menyerah begitu saja, yang seringkali merupakan tujuannya.

Pola Gelap dan AI: Tanda-Tanda Peringatan Utama
• Antarmuka yang menyembunyikan atau mengaburkan tombol opt-out dan berhenti berlangganan • "Konfirmasi yang memalukan" (misalnya, "Tidak, terima kasih, saya lebih suka membayar harga penuh") • Tombol atau opsi yang diberi label menyesatkan • Pengujian A/B waktu nyata yang meningkatkan konversi tetapi mengurangi kejelasan
Beberapa negara bagian, termasuk California dan Colorado, telah mengadopsi larangan pola gelap dalam undang-undang konsumen dan terkait AI mereka.

Chatbot, Peniruan Identitas, dan Interaksi AI yang Menyesatkan

Seiring dengan semakin sulitnya membedakan AI percakapan dari percakapan manusia sungguhan, regulator berfokus pada memastikan bahwa identitas pengguna diungkapkan ketika mereka berinteraksi dengan mesin, terutama dalam konteks komersial atau konsultasi. Misalnya, sebuah chatbot menawarkan panduan dalam kategori seperti hukum, kesehatan, spiritual, atau keuangan, tetapi konsumen secara wajar percaya bahwa mereka berinteraksi

dengan seorang ahli manusia. Dalam hal ini, kurangnya pengungkapan dapat dianggap sebagai kelalaian material berdasarkan standar UDAP.

Suara yang dihasilkan AI dan deepfake semakin banyak digunakan dalam penipuan atau dukungan yang menyesatkan. Dengan atau tanpa niat jahat, meniru seseorang untuk keuntungan komersial atau pribadi dapat melanggar undang-undang penipuan konsumen, undang-undang hak publisitas, atau panduan dari FTC mengenai dukungan.

Ulasan Palsu, Dukungan yang Dihasilkan AI, dan Penipuan Influencer

Kepercayaan konsumen terhadap ulasan dan testimoni merupakan aspek mendasar dari perdagangan online. Namun, semakin banyak situs perdagangan online memiliki ulasan yang reputasinya dipertanyakan, baik yang dihasilkan oleh AI atau ulasan untuk secara artifisial meningkatkan kualitas produk. AI sekarang memungkinkan:

- ✓ Ulasan palsu yang dihasilkan secara massal
- ✓ Umpan balik positif yang dihasilkan secara otomatis untuk produk sendiri
- ✓ Influencer deepfake yang mempromosikan barang atau jasa tanpa pengungkapan

Pada tahun 2023, FTC memperbarui Panduan Dukungan dan mengusulkan aturan yang mewajibkan pengungkapan testimoni yang dihasilkan AI, larangan ulasan palsu atau tidak terverifikasi, dan akuntabilitas bagi bisnis yang mendapat manfaat dari ulasan yang menipu, bahkan jika dihasilkan oleh afiliasi atau vendor pihak ketiga. Ulasan yang dihasilkan AI yang tampak autentik tetapi sebenarnya palsu melanggar standar "konsumen yang wajar" dan kemungkinan besar akan dianggap menipu, yang mengakibatkan tindakan hukum dan sanksi.

Risiko yang Muncul: Deepfake, Disinformasi, dan Manipulasi Emosional

Kemampuan AI untuk menghasilkan gambar, video, dan audio yang realistis membuka babak baru penipuan konsumen dan bisnis, terutama karena alat-alat tersebut menjadi lebih mudah diakses dan lebih ampuh. Saat buku ini ditulis, penulis mencatat bahwa alat-alat yang tersedia untuk membuat deepfake tersedia secara luas dan menghasilkan konten yang sangat realistis.

Deepfake dapat digunakan untuk memalsukan dukungan selebriti, membuat bukti palsu tentang peristiwa atau pernyataan, memalsukan identitas pemerintah, meniru suara atau wajah untuk penipuan, dan menyebarkan informasi politik atau keuangan palsu. Pada Mei 2025, Amerika Serikat memberlakukan Undang-Undang Take It Down, yang mengkriminalisasi distribusi gambar intim tanpa persetujuan, termasuk deepfake AI. Selain itu, Undang-Undang Keamanan Online Inggris menganggap berbagi konten deepfake eksplisit tanpa persetujuan sebagai tindak pidana. Deepfake juga dapat dituntut berdasarkan:

- ➢ Hukum praktik yang tidak adil atau menipu, jika digunakan untuk pemasaran atau keuntungan;
- ➢ Hukum hak publisitas, jika kemiripan disalahgunakan;
- ➢ Undang-undang penipuan kawat, jika digunakan untuk penipuan peniruan identitas.

Beberapa negara bagian (misalnya, California, Texas, dan New York) telah memberlakukan undang-undang yang menargetkan penggunaan deepfake yang berbahaya dalam konten politik atau pornografi, dan kemungkinan lebih banyak negara bagian akan mengikutinya.

Tanda-Tanda Bahaya Deepfake
❖ Dukungan, pengakuan, dan arahan dengan kualitas yang tidak realistis ❖ Video yang dibagikan tanpa metadata atau sumber asli ❖ Suara sintetis yang digunakan dalam panggilan layanan pelanggan atau dukungan tanpa pemberitahuan ❖ Pernyataan yang tidak sesuai dengan karakter—seseorang mengatakan hal-hal yang bertentangan dengan pandangan atau gaya biasanya

Disinformasi AI dalam Kesehatan, Keuangan, dan Keselamatan

Konsumen dapat menderita kerugian ketika model AI generatif memberikan nasihat kesehatan yang tidak akurat, panduan keuangan yang salah, dan instruksi keselamatan yang keliru. Bahkan tanpa niat jahat atau menyesatkan, perusahaan dapat dimintai pertanggungjawaban ketika AI dipasarkan sebagai sesuatu yang dapat dipercaya atau berwibawa, konsumen mengandalkan informasi tersebut dalam penelitian dan pengambilan keputusan mereka, dan jika kerugian tersebut dapat diperkirakan dan dicegah.

Berikut contoh nyata: meskipun ChatGPT menyertakan penafian bahwa mereka bukan penyedia layanan medis, banyak yang mengandalkan outputnya untuk mendiagnosis gejala atau interaksi obat. Ketergantungan yang tidak tepat dapat menyebabkan klaim hukum.

Manipulasi Emosional dan Populasi Rentan

Alat AI yang berinteraksi dengan konsumen yang rentan secara emosional, seperti anak-anak, lansia, atau mereka yang mengalami duka cita atau masalah kesehatan mental, mungkin akan mendapat pengawasan ketat. Kekhawatiran ini sangat relevan dalam chatbot terapeutik, aplikasi pendamping duka cita berbasis AI, dan konten yang dihasilkan AI untuk anak-anak.

Jika alat-alat ini salah menggambarkan kemampuannya atau mengeksploitasi keadaan emosional untuk keterlibatan, keuntungan, atau penipuan, mereka dapat melanggar standar ketidakadilan berdasarkan Bagian 5 dari Undang-Undang FTC.

Hubungan Romantis dengan AI

Alat AI semakin banyak digunakan untuk persahabatan romantis. Sebuah survei luas di Amerika Serikat oleh Wheatley Institute menemukan bahwa sekitar 19% orang dewasa muda berusia 18–30 tahun telah berinteraksi dengan chatbot AI romantis. Lebih lanjut, 31% pria muda dan 23% wanita muda telah melaporkan interaksi tersebut. Sebuah studi yang didukung YouGov mengungkapkan bahwa 7% orang dewasa yang belum menikah mengaku terbuka untuk, atau saat ini terlibat dalam, hubungan romantis dengan pasangan AI.

Institut Studi Keluarga melakukan penelitian yang mengungkapkan bahwa satu dari empat orang dewasa muda percaya bahwa pasangan AI dapat menggantikan romansa kehidupan nyata. Peraturan FTC kemungkinan berlaku untuk layanan romansa AI jika layanan tersebut salah menggambarkan kemampuannya atau gagal melindungi informasi pribadi.

Strategi Mitigasi Risiko dan Kepatuhan Perlindungan Konsumen

Menghindari hasil AI yang tidak adil atau menipu bukan hanya keharusan hukum; Hal ini sangat penting untuk membangun kepercayaan dan meminimalkan risiko reputasi dan

regulasi. Bagian ini menguraikan bagaimana organisasi dapat mengoperasionalkan prinsip-prinsip perlindungan konsumen melalui dokumentasi, desain, dan akuntabilitas.

Desain untuk Transparansi dan Keterbukaan

Konsumen harus secara eksplisit memahami kapan AI terlibat dalam pengambilan keputusan atau interaksi. Organisasi harus memastikan bahwa transparansi ini merupakan bagian dari desain setiap sistem AI. Transparansi ini dapat dicapai melalui:

- Pengungkapan yang jelas bahwa pengguna berinteraksi dengan AI (misalnya, “Saya adalah asisten AI, bukan perwakilan manusia”)
- Penjelasan yang mudah diakses (tidak tersembunyi dalam syarat dan ketentuan layanan) tentang bagaimana keputusan dibuat (misalnya, penolakan kredit, penetapan harga)
- Penafian kontekstual bahwa respons AI bukanlah nasihat profesional dalam perawatan kesehatan, keuangan, atau hukum
- Pernyataan keterbatasan model, yang menunjukkan di mana AI mungkin tidak lengkap, tidak pasti, atau rentan terhadap kesalahan
- Sumber daya bantuan yang mudah ditemukan atau jalur eskalasi ke perwakilan manusia

Pengungkapan “Anda berinteraksi dengan AI” ini harus menonjol (tidak tersembunyi dalam syarat dan ketentuan layanan). Pengungkapan ini harus diungkapkan dalam bahasa yang mudah dipahami dan diberikan saat AI digunakan, bukan hanya pada saat layanan atau orientasi sesi.

Ketika Google memperkenalkan asisten suara AI Google Duplex yang mampu melakukan panggilan telepon, mereka secara eksplisit memprogram alat tersebut untuk mengidentifikasi dirinya sebagai AI selama panggilan, sebuah langkah transparansi yang dipuji oleh beberapa regulator dan pendukung. Sebagai catatan, Google menutup Duplex pada tahun 2022 dan sekarang menawarkan panggilan keluar melalui produk Pencarian mereka.

Membuktikan Klaim AI

Klaim pemasaran oleh organisasi tentang sistem AI mereka harus akurat dan dapat dibuktikan. Pelanggaran umum meliputi:

- Melebih-lebihkan kemampuan (misalnya, "akurasi terjamin" atau "penilaian setara manusia")
- Dukungan atau testimoni yang menyesatkan
- Menyiratkan bahwa AI lebih aman atau lebih objektif tanpa memberikan bukti

Klaim kuantitatif, seperti tingkat akurasi atau penghematan biaya, harus didukung oleh data yang kuat dan pengujian yang dapat direproduksi.

Seringkali, organisasi akan menambahkan AI ke layanan yang sudah ada dan kemudian membanggakan kemampuan "bertenaga AI" mereka yang unggul. Karena begitu banyak organisasi yang kekurangan tata kelola AI yang efektif, klaim ini terkadang hanyalah angan-angan belaka.

Daftar Periksa Kepatuhan Pemasaran AI

Saat membuat konten pemasaran terbaru yang mempromosikan kemampuan AI, pemasar harus memeriksa bahwa:

- Semua klaim kinerja AI didukung oleh bukti kuantitatif atau kualitatif.
- Lingkup peran AI dinyatakan dengan jelas (dibantu manusia vs. sepenuhnya otomatis).
- Penafian terlihat di tempat yang diperlukan.
- Keterbatasan alat pihak ketiga diungkapkan.
- Testimoni otentik dan diberi label dengan tepat.

Memantau Pergeseran dan Kerugian Algoritma

Sistem AI dapat mengalami penurunan kualitas seiring waktu, beradaptasi dengan data baru dengan cara yang tidak terduga, atau berinteraksi dengan perilaku pengguna dalam pola yang muncul dan berisiko. Terlebih lagi, meskipun model AI tidak mengalami penurunan kualitas seiring waktu, relevansinya tetap dapat terganggu: model AI tahun lalu mungkin tidak lagi bermanfaat bagi bisnis tahun ini. Penurunan kualitas ini dan bentuk penurunan kualitas lainnya menjadikan pemantauan berkelanjutan sangat penting. Praktik terbaik meliputi:

- Audit kinerja dan keadilan rutin
- Pengujian tim merah untuk mendeteksi perilaku yang menyalahgunakan atau manipulatif
- Pencatatan dan peninjauan keluhan pengguna
- Pemantauan terhadap positif palsu, halusinasi, atau respons yang tidak aman
- Pemantauan keamanan untuk upaya injeksi cepat, peracunan data, atau eksfiltrasi model

Sistem AI harus memiliki sakelar pemutus, pengamanan, pemantauan peristiwa dan insiden, serta jalur eskalasi untuk memungkinkan intervensi manusia.

Menerapkan Kebijakan Tata Kelola AI

Seperti yang telah berulang kali dinyatakan dalam buku ini, kebijakan adalah "hukum" dalam organisasi. Saat membangun proses tata kelola AI, proses tersebut harus didukung oleh kebijakan. Untuk memastikan perlindungan konsumen tertanam dalam desain, pengembangan, dan operasi AI, organisasi harus menetapkan:

- ✓ Komite peninjauan produk AI yang mencakup masukan hukum, UX, keamanan siber, privasi, dan etika
- ✓ Daftar risiko AI lintas fungsi
- ✓ Kartu model atau lembar data terdokumentasi yang menjelaskan tujuan, keterbatasan, dan risiko yang diketahui
- ✓ Rencana respons insiden untuk keluhan konsumen terkait AI

Tata kelola harus mencakup seluruh siklus hidup AI, mulai dari pengumpulan data dan pelatihan model hingga penerapan dan penghentian.

Kontrol Risiko Vendor dan Pihak Ketiga

Banyak layanan AI yang berhadapan langsung dengan konsumen bergantung pada alat, API, dan kumpulan data pihak ketiga. Organisasi tetap bertanggung jawab secara hukum atas bagaimana komponen-komponen ini berfungsi, termasuk peristiwa operasional di luar kendali mereka.

Untuk mengelola risiko ini, organisasi harus memeriksa vendor untuk riwayat kepatuhan dan praktik audit, memasukkan kewajiban perlindungan konsumen dalam kontrak, mensyaratkan ketentuan hak untuk audit, dan menetapkan ketentuan pemberitahuan insiden jika terjadi kegagalan model atau kerugian pengguna. Persyaratan ini merupakan bagian dari harapan inti dalam praktik manajemen risiko pihak ketiga.

Pendekatan Proaktif terhadap AI dan Perlindungan Konsumen

AI mengubah cara konsumen mengakses layanan, membuat keputusan, dan berinteraksi dengan dunia, tetapi juga memperkenalkan jalan baru untuk manipulasi, penipuan, dan kerugian. Hukum perlindungan konsumen menyediakan kerangka kerja yang fleksibel dan netral terhadap teknologi untuk meminta pertanggungjawaban perusahaan ketika AI menyebabkan hasil yang tidak adil atau menyesatkan.

Saat mengidentifikasi hukum dan peraturan yang berlaku dalam AI dan privasi, organisasi harus mendokumentasikan pendekatan mereka terhadap kepatuhan terhadap peraturan tersebut. Namun, karena banyak aspek AI yang belum diatur, organisasi harus berhati-hati dan berasumsi bahwa kekosongan peraturan tersebut pada akhirnya akan terisi. Oleh karena itu, organisasi harus secara proaktif:

- Merancang transparansi dan keadilan ke dalam produk dan kemampuan AI sejak awal
- Melakukan audit untuk mengetahui adanya kerugian dan perilaku yang tidak diinginkan
- Menghindari janji berlebihan terhadap kemampuan AI
- Mengungkapkan kapan pengguna berinteraksi dengan mesin
- Memantau bagaimana konsumen di dunia nyata mengalami alat AI

Kepercayaan pada AI akan lebih bergantung pada transparansi, integritas hukum, dan tanggung jawab etis organisasi terkait sistem AI dan kurang bergantung pada kinerja teknis. Ketika organisasi mengikuti prinsip-prinsip ini, mereka cenderung tidak perlu melakukan penyesuaian ulang pada sistem AI mereka ketika peraturan baru diberlakukan.

Tanda-Tanda Peringatan AI untuk Tim Produk

Bendera Merah	**Mengapa Ini Berisiko**
Klaim AI "100% akurat"	Bisa jadi menyesatkan kecuali dapat dibuktikan.
Chatbot AI memberikan nasihat medis atau hukum.	Dapat menyesatkan konsumen dalam mengambil keputusan yang tidak aman.
Tidak ada pengungkapan saat pengguna berinteraksi dengan AI.	Melanggar ekspektasi pengungkapan material
Mengoptimalkan UI untuk keterlibatan daripada kejelasan.	Dapat memicu penerapan pola gelap
Ulasan atau dukungan yang dihasilkan oleh AI	Kemungkinan menyesatkan kecuali diungkapkan secara jelas.

Klaim yang didasarkan pada data pelatihan yang sudah usang atau bias.	Dapat menyebabkan hasil yang tidak adil tanpa jalan keluar.

Tanda-tanda peringatan ini seharusnya memicu tinjauan internal dan pemeriksaan hukum sebelum diterapkan.

5.4 BAGAIMANA HUKUM TANGGUNG JAWAB PRODUK BERLAKU UNTUK AI

Sistem AI semakin banyak tertanam dalam produk yang berinteraksi dengan dunia fisik: mobil, perangkat medis, drone, peralatan rumah tangga, dan robot industri dan rumah tangga. Ketika sistem ini mengalami malfungsi atau berperilaku dengan cara yang tidak terduga, konsekuensinya dapat menjadi bencana dan bahkan mengancam jiwa. Hukum tanggung jawab produk menyediakan kerangka hukum untuk menentukan siapa yang bertanggung jawab ketika produk menyebabkan kerugian. Namun, kategori tradisional cacat desain, manufaktur, dan peringatan dikembangkan pada era sebelum sistem otonom dan adaptif. Bagian ini mengeksplorasi bagaimana hukum tanggung jawab produk berlaku untuk AI dan menguraikan tantangan hukum yang muncul dalam menetapkan tanggung jawab atas kerugian yang disebabkan oleh AI.

Tanggung Jawab Produk dan Penerapannya pada AI

Tanggung jawab produk adalah cabang hukum perdata yang menuntut pertanggungjawaban produsen, perancang, dan penjual atas produksi dan penjualan produk yang cacat. Tanggung jawab produk sudah mapan untuk barang-barang berwujud; namun, penerapannya pada sistem berbasis perangkat lunak, otonom, atau pembelajaran tidaklah mudah.

Teori Tanggung Jawab Produk

Klaim tanggung jawab produk di Amerika Serikat biasanya didasarkan pada satu atau lebih hal berikut:

- Cacat desain: Suatu produk diproduksi sesuai dengan yang dimaksudkan, tetapi desain itu sendiri secara tidak wajar berbahaya atau berisiko.
- Cacat manufaktur: Produk menyimpang dari desain yang dimaksudkan karena kesalahan dalam pembuatan atau perakitan.
- Kegagalan untuk memperingatkan (cacat pemasaran): Produk tidak memiliki instruksi atau peringatan yang memadai tentang risiko.

Hal ini diterapkan melalui:

- Tanggung jawab mutlak: Penggugat tidak perlu membuktikan kelalaian, hanya bahwa produk tersebut cacat dan menyebabkan kerugian.
- Kelalaian: Penggugat harus menunjukkan bahwa tergugat gagal bertindak dengan kehati-hatian yang wajar dalam desain, pembuatan, pengujian, atau pengawasan.
- Pelanggaran garansi: Berdasarkan janji kontraktual tentang keamanan atau kinerja produk.

Tanggung Jawab Mutlak vs. Kelalaian		
Fitur	**Tanggung Jawab Ketat**	**Kelalaian**
Standar	Produk cacat menyebabkan kerugian	Kegagalan untuk bertindak dengan kehati-hatian yang wajar
Maksud	Tidak diperlukan	Diperlukan (kegagalan untuk mencegah kerugian yang dapat diperkirakan)
Umum Kasus Penggunaan	Cacat desain dan produksi	Pengujian, pengawasan, atau peringatan yang tidak memadai
Contoh AI	Penyedot debu otonom meledak karena suatu kerusakan.	Keputusan AI menimbulkan kerugian karena pengawasan yang buruk.

Apa yang Dimaksud dengan "Cacat" pada Produk yang Digerakkan AI?

Dalam hukum tanggung jawab produk tradisional, cacat mengacu pada kekurangan yang membuat suatu produk menjadi sangat berbahaya jika digunakan sesuai petunjuk. Dengan sistem AI, konsep cacat menjadi lebih kabur karena beberapa alasan. Pertama, perilaku sistem AI bersifat dinamis dan sensitif terhadap konteks. Selain itu, model AI dapat berubah (berubah setelah penerapan awal), dan hasilnya bersifat probabilistik, bukan deterministik.

Para ahli hukum dan pengadilan masih bergulat dengan pertanyaan-pertanyaan seperti:

- Apakah cacat AI merupakan kesalahan pengkodean, kekurangan data, atau keputusan yang tidak terduga?
- Siapa yang bertanggung jawab ketika suatu produk berkinerja berbeda berdasarkan perilaku pengguna atau input data?

Mengenai tanggung jawab produk, produsen sering kali bertanggung jawab atas penyalahgunaan yang dapat diprediksi, tetapi dengan AI, "dapat diprediksi" lebih sulit didefinisikan karena perilaku yang muncul. Dapat diprediksi menjadi medan pertempuran hukum utama.

Pada tahun 2018, sebuah kendaraan Uber otonom menabrak dan membunuh seorang pejalan kaki di Arizona selama fase pengujian. Investigasi menemukan bahwa perangkat lunak gagal mengklasifikasikan orang tersebut sebagai bahaya karena desain deteksi objek yang buruk. Apakah cacat tersebut terletak pada algoritma, konfigurasi sensor, atau protokol pengujian? Ambiguitas ini menggambarkan betapa sulitnya untuk menentukan dan membuktikan "cacat" dalam sistem yang membuat keputusan independen.

Pada tahun 2025, Hertz, Thrifty, Avis, Sixt, dan perusahaan penyewaan mobil lainnya menggunakan sistem pemindaian AI untuk memeriksa kerusakan pada mobil sewaan yang dikembalikan. Kemudian keluhan mulai berdatangan—pelanggan dikenakan biaya untuk kerusakan yang, dalam kebanyakan kasus, hanyalah goresan kecil atau bahkan tidak terlihat sama sekali. Namun, bagian terburuknya adalah tidak ada pengawasan manusia atau proses banding yang tersedia. Pada saat penulisan buku ini, ini adalah cerita yang sedang berkembang.

Memperluas Definisi "Produk" untuk Mencakup Perangkat Lunak dan AI

Di bawah kerangka hukum tradisional, pengadilan enggan memperlakukan perangkat lunak murni sebagai produk karena perangkat lunak dipandang sebagai layanan (bukan barang berwujud), perjanjian lisensi (bukan penjualan) mengatur transaksi, dan perilaku perangkat lunak tidak dapat diprediksi sepenuhnya.

Seiring semakin banyaknya kasus pengadilan tentang tanggung jawab produk teknologi tinggi yang disidangkan, dan seiring semakin banyaknya pengacara dan hakim yang mempelajari tentang AI, posisi hukum pun bergeser. Pengadilan semakin bersedia memperlakukan perangkat lunak tertanam dan AI sebagai bagian dari sistem produk, terutama ketika perangkat lunak tersebut secara langsung mengontrol perangkat keras fisik, fungsi perangkat lunak tersebut penting untuk keselamatan, dan AI melakukan tugas-tugas yang menggantikan penilaian manusia.

Restatement (Third) of Torts: Tanggung Jawab Produk

Restatement (Third) yang berpengaruh mendefinisikan produk sebagai “harta benda pribadi berwujud yang didistribusikan secara komersial untuk digunakan atau dikonsumsi.” Meskipun perangkat lunak tidak selalu termasuk, perangkat lunak tertanam yang mengoperasikan produk fisik (misalnya, autopilot, pompa insulin, drone) mungkin memenuhi syarat.

Cacat Desain pada Sistem AI

Klaim cacat desain umumnya menyatakan bahwa desain suatu produk secara inheren cacat dan menciptakan risiko yang tidak perlu bagi penggunanya. Dalam konteks AI, klaim ini berfokus pada arsitektur model, pemilihan data pelatihan, asumsi risiko, atau kegagalan untuk memperhitungkan penyalahgunaan yang dapat diperkirakan.

Penyeimbangan Risiko-Manfaat dan “Desain Alternatif yang Wajar”

Di Amerika Serikat, pengadilan sering menerapkan aturan yang disebut uji risiko-manfaat yang menanyakan, “Apakah risiko bahaya lebih besar daripada manfaat produk, dan apakah desain alternatif yang wajar dapat mengurangi risiko tanpa mengorbankan fungsionalitas?”

Dalam konteks AI, penggugat dalam gugatan tanggung jawab produk dapat menyatakan bahwa sistem AI menggunakan data pelatihan yang tidak memadai atau tidak tepat, bahwa arsitektur model yang lebih aman tersedia, desain gagal menyertakan pengawasan manusia, atau perilaku default yang lebih aman dimungkinkan.

Dalam contoh hipotetis, asisten bedah robotik membuat sayatan yang tidak tepat karena gambar yang salah klasifikasi dari sistem penglihatan AI-nya, dan pasien menderita kerugian. Jika algoritma klasifikasi yang lebih aman tersedia dan diabaikan, atau verifikasi manusia dapat mencegah kesalahan tersebut, klaim cacat desain mungkin berhasil.

Cacat Manufaktur pada Produk Terintegrasi AI

Cacat manufaktur terjadi ketika produk akhir berbeda dari desain yang dimaksudkan karena kesalahan manufaktur. Klaim ini umumnya lebih mudah dibuktikan daripada cacat desain karena melibatkan penyimpangan yang jelas dari desainnya.

Cacat manufaktur dalam konteks AI meliputi:

- Sensor atau kamera yang rusak yang dipasang di kendaraan otonom
- Firmware atau penyebaran perangkat lunak yang rusak
- Kegagalan perangkat keras yang mengganggu persepsi atau kontrol AI
- Kesalahan integrasi antara komponen AI dan sistem mekanis
- Masalah kompatibilitas dengan komponen yang diproduksi oleh pihak ketiga

AI sebagai Rakitan Bagian yang Saling Bergantung

Karena produk yang didukung AI bergantung pada perangkat keras, perangkat lunak, konektivitas, dan input data, bahkan variasi kecil dalam satu komponen dapat menyebabkan perilaku yang tidak dapat diprediksi. Dapat dikatakan bahwa produsen AI bukanlah produsen tetapi integrator, yang merakit komponen dari beberapa sumber, menghasilkan sistem AI akhir.

Robot industri yang dipandu oleh AI dirancang untuk berhenti ketika bertemu dengan manusia. Selama proses produksi, sensor jarak yang tidak dikalibrasi dengan benar dipasang, dan robot gagal berhenti, sehingga melukai seorang pekerja. Meskipun perangkat lunak AI bekerja dengan benar, penyimpangan perangkat keras tersebut merupakan cacat produksi. Jika hal ini berujung pada proses hukum, pengadilan atau arbiter kemungkinan akan menetapkan tanggung jawab mutlak jika cacat tersebut dapat ditelusuri ke proses manufaktur, produk akhir digunakan sesuai tujuan, dan kerugian tersebut dapat diperkirakan dan secara langsung disebabkan oleh cacat tersebut.

Tantangan dalam Mendeteksi Cacat Manufaktur AI

Karena sistem AI bersifat non-deterministik, membedakan variasi normal dari cacat sebenarnya dapat menjadi tantangan, yang mempersulit beban pembuktian penggugat dalam klaim tanggung jawab tradisional. Isu-isu utama dalam tantangan hukum meliputi kurangnya transparansi dalam pengambilan keputusan sistem AI, pergeseran model, dan kompleksitas rantai pasokan.

Kesulitan-kesulitan ini telah mendorong diskusi tentang pergeseran beban pembuktian, yang mengharuskan produsen untuk membuktikan bahwa suatu sistem tidak cacat setelah terjadi kerugian. Pergeseran tersebut tetap kontroversial dan bergantung pada yurisdiksi.

Bayangkan, jika Anda mau, sebuah insiden antara kendaraan otonom yang menabrak kanguru yang membawa tangga. Pemilik kanguru mungkin menggugat berdasarkan cacat manufaktur. Bagaimana produsen kendaraan tersebut diharuskan untuk membuktikan bahwa kendaraannya bebas dari cacat, termasuk kasus penggunaan yang paling tidak mungkin dalam pengujian keselamatan?

Kegagalan untuk Memperingatkan: Mengungkapkan Risiko AI

Hukum tanggung jawab produk juga mengharuskan produsen produk untuk memperingatkan konsumen tentang bahaya yang tidak jelas terkait dengan penggunaan

produk. Kewajiban ini bersifat berkelanjutan: kewajiban ini berlaku pada saat penjualan dan dapat diperpanjang seiring dengan perkembangan risiko setelah implementasi.

Para pembuat sistem AI konsumen mungkin perlu mempertimbangkan untuk memberi tahu pelanggan ketika:

- Instruksi pengguna menghilangkan keterbatasan akurasi AI yang diketahui
- Penafian produk tersembunyi atau tidak ditulis dalam bahasa yang mudah dipahami
- Pengguna tidak menyadari bahwa AI mengendalikan atau memengaruhi hasil
- Pengguna tidak diberi tahu tentang kasus-kasus ekstrem atau kondisi risiko

Mengingat kecenderungan perusahaan rintisan teknologi tinggi untuk fokus terlebih dahulu pada waktu pemasaran dan kemudian (jika ada) pada kepatuhan terhadap hukum yang berlaku, kita mungkin akan melihat musim tindakan penegakan hukum atas cacat produk terhadap vendor sistem AI.

Antarmuka Manusia-Mesin dan Kesalahpahaman Pengguna

Pengguna mungkin menjadi terlalu bergantung pada sistem AI, dengan asumsi bahwa sistem tersebut akurat dan lebih mampu daripada yang sebenarnya. Risiko ini meningkat ketika sistem AI dipasarkan sebagai "otonom," "belajar mandiri," atau "aman," atau ketika UI tidak memberikan indikasi ketidakpastian atau kesalahan, atau ketika panduan pengguna tidak memadai untuk penggunaan di dunia nyata.

Meskipun fitur "Autopilot" dan "Full Self-Driving" Tesla memerlukan pengawasan pengemudi yang konstan, beberapa konsumen telah salah memahami kemampuannya. Kebingungan ini telah menyebabkan litigasi dan pengawasan regulasi, dengan para kritikus berpendapat bahwa penamaan merek itu sendiri merupakan kegagalan untuk memberikan peringatan. Pada pertengahan tahun 2025, juri federal memberikan ganti rugi lebih dari $240 juta kepada keluarga pemilik Tesla yang meninggal dalam kecelakaan terkait Tesla Autopilot pada tahun 2019.

Memperbarui Peringatan Pasca-Peluncuran

Karena sistem AI berkembang dan risiko baru dapat muncul setelah peluncuran, vendor sistem AI mungkin memiliki kewajiban berkelanjutan untuk memberikan peringatan, termasuk pemberitahuan pembaruan perangkat lunak, peringatan keselamatan, pengumuman penarikan kembali, dan pengungkapan publik tentang masalah yang diketahui. Pengadilan belum secara pasti menetapkan sejauh mana kewajiban ini berlaku, terutama untuk sistem AI yang belajar atau beradaptasi di lapangan.

Preemption dan Ketegangan Regulasi

Dalam industri yang diatur seperti perawatan kesehatan, otomotif, dan penerbangan, perusahaan dapat berpendapat bahwa peraturan federal mengesampingkan klaim tanggung jawab produk tingkat negara bagian. Misalnya, ketika FDA (*Food and Drug Administration*) menyetujui alat diagnostik medis bertenaga AI untuk penggunaan komersial, negara bagian AS seharusnya tidak diizinkan untuk menegakkan hukum tanggung jawab produk negara bagian. Dan demikian pula, produsen mobil yang sistem pengemudian otonomnya telah disertifikasi oleh NHTSA (*National Highway Traffic Safety Administration*) juga akan berpendapat bahwa negara bagian AS seharusnya tidak menegakkan hukum tanggung jawab produk negara bagian.

Seiring munculnya kasus hukum baru, pengadilan perlu menyeimbangkan peran pengawasan federal dengan hak konsumen untuk mendapatkan ganti rugi atas cedera.

Pencegahan dan AI	
Mengeklaim	**Potensi Pencegahan**
Cacat desain pada perangkat medis berbasis AI.	Dapat didahului oleh persetujuan FDA.
Kegagalan memberikan peringatan tentang AI pada drone	Kemungkinan kecil untuk dicegah
Pembaruan pasca-penjualan menyebabkan kerugian	Kecil kemungkinannya untuk dilindungi
Meskipun preemption dapat membatasi tanggung jawab, pengadilan cenderung lebih mengutamakan peringatan dan hak konsumen ketika terdapat ambiguitas.	

Ketidakpastian Hukum yang Berlanjut

Karena cedera produk yang disebabkan oleh AI relatif baru, hukum kasus masih terbatas tetapi terus berkembang. Pengadilan belum menetapkan preseden yang tegas tentang evaluasi kausalitas dalam sistem yang bergantung pada probabilitas, siapa yang bertanggung jawab ketika sistem otonom membuat keputusan yang merugikan, dan apakah AI dapat dianggap sebagai produk.

Tanggung Jawab Multipihak dan Rantai Pasokan AI

Rantai pasokan untuk produk AI cukup kompleks dan berbeda secara signifikan dari rantai pasokan produk konsumen tradisional. Tanggung jawab produk secara tradisional memungkinkan tuntutan hukum terhadap perancang produk, produsen, distributor, dan penjual. Namun, rantai pasokan sistem AI sangat berbeda dan terdiri dari vendor perangkat keras, penyedia layanan cloud, pengembang model, broker data, dan integrator sistem. Dalam tindakan hukum terkait tanggung jawab, aktor rantai pasokan sistem AI cenderung menyangkal tanggung jawab dan menyalahkan pihak lain.

Berikut adalah contoh hipotetis: pembaruan firmware dalam sistem rumah pintar bertenaga AI mengalami malfungsi dan menonaktifkan detektor asap, yang berkontribusi pada kebakaran rumah. Tanggung jawab dapat diperselisihkan antara produsen perangkat pintar, penyedia AI berbasis cloud, dan vendor perangkat lunak yang membuat dan meluncurkan pembaruan tersebut. Contoh ini mengingatkan penulis pada situasi di mana pemilik mobil mengalami masalah pengoperasian, di mana vendor roda menyalahkan ban, dan vendor ban menyalahkan roda, sehingga pemilik tidak mendapatkan solusi. Ambiguitas hukum tentang tanggung jawab bersama dan tanggung jawab secara terpisah dalam konteks AI membuat proses hukum dan litigasi menjadi lebih kompleks.

Pendekatan Uni Eropa terhadap AI dan Tanggung Jawab Produk

Uni Eropa sedang berupaya memodernisasi rezim tanggung jawab produknya untuk mencakup risiko terkait AI. Pada tahun 2022, Komisi Eropa mengusulkan revisi Arahan Tanggung Jawab Produk untuk mencakup perangkat lunak dan pembaruan digital, bersama

dengan asumsi cacat dalam kasus AI berisiko tinggi (misalnya, kendaraan otonom), serta perluasan tanggung jawab kepada pengembang dan penyebar, bukan hanya produsen.

Meskipun tidak berlaku secara langsung di Amerika Serikat, proposal ini dapat memengaruhi peraturan AS, litigasi, desain produk, dan strategi kepatuhan secara keseluruhan.

Strategi Mitigasi Risiko Proaktif untuk Tanggung Jawab Produk AI

Karena litigasi tanggung jawab produk bersifat reaktif (dipicu hanya setelah kerugian terjadi), organisasi yang menerapkan AI dalam produk konsumen atau industri harus mengadopsi strategi risiko preventif. Ini termasuk kontrol rekayasa, dokumentasi, tata kelola lintas fungsi, dan validasi eksternal. Dengan kata lain, produk AI harus aman, terjamin, transparan (di mana kemampuan menjelaskan menjadi faktor), dan sesuai sejak tahap desain. Bagian ini mencakup beberapa langkah proaktif yang dapat dipertimbangkan oleh organisasi dalam rantai pasokan sistem AI.

Melakukan Pengujian Keamanan Khusus AI

Perluas pengujian produk AI untuk mempertimbangkan:

- ✓ Perilaku kasus ekstrem di lingkungan non-deterministik
- ✓ Audit bias dan keadilan untuk output yang memengaruhi hasil fisik
- ✓ Pengujian simulasi untuk peristiwa langka tetapi dahsyat
- ✓ Risiko interaksi manusia-AI, termasuk ketergantungan berlebihan dan kebingungan
- ✓ Risiko kerentanan, termasuk patch yang merugikan, injeksi cepat, dan peracunan data

Praktik penting dalam pengujian keamanan sistem AI meliputi:

- ➢ Menggunakan kembaran digital atau lingkungan sandbox untuk mensimulasikan penggunaan dunia nyata
- ➢ Pengujian stres model AI menggunakan input yang tidak dapat diprediksi (mirip dengan fuzzing dalam keamanan siber)
- ➢ Memvalidasi keputusan model terhadap tolok ukur dan pedoman peraturan

Sebelum penyebaran, sistem drone otonom yang digerakkan AI diuji terhadap pemalsuan GPS, pergeseran angin, penyeberangan pejalan kaki, dan kegagalan sensor. Simulasi ini mengungkapkan pola penurunan ketinggian yang tidak terduga—memungkinkan koreksi sebelum terjadi kerusakan di dunia nyata.

Pertahankan Keterlacakan di Seluruh Siklus Hidup AI

Klaim tanggung jawab produk sering kali bergantung pada pembuktian (atau penyangkalan) asal usul cacat produk. Tata kelola AI untuk setiap organisasi dalam rantai pasokan AI harus memastikan keterlacakan asal usul dan pengarsipan data pelatihan, versi model dan log pembaruan, hasil pengujian dan artefak validasi, serta komitmen kode sumber dan catatan penyebaran. Dengan kata lain, setiap artefak dalam siklus hidup pengembangan komponen harus dicatat dan diarsipkan.

Organisasi rantai pasokan AI harus memelihara daftar komponen sistem AI (AI-SBOM) yang mendokumentasikan ketergantungan dan subkomponen dari sistem AI. Rekomendasi ini mirip dengan daftar komponen perangkat lunak (SBOM), inventaris komponen yang digunakan untuk membangun perangkat lunak, yang semakin umum di industri perangkat lunak.

Apa yang Harus Dilacak dalam AI-SBOM

Para vendor dalam rantai pasokan sistem AI dapat mulai membangun AI-SBOM mereka untuk mencakup:

Elemen	Tujuan
Versi model dan ID	Identifikasi versi mana yang menyebabkan kerusakan
Sumber kumpulan data	Mengungkapkan bias atau masalah perizinan
Langkah-langkah pra-pemrosesan	Jelaskan perilaku model yang tidak terduga
Log firmware/perangkat lunak	Validasi konfigurasi pada saat terjadi kesalahan.
Perbarui stempel waktu	Tentukan apakah perubahan terkini merupakan penyebab.

Organisasi yang mengevaluasi sistem AI harus mempertimbangkan apakah vendor sistem AI memiliki AI-SBOM (AI-SBOM). Tingkat dokumentasi ini dapat mengurangi tanggung jawab dengan menunjukkan ketelitian dan pengendalian yang semestinya.

Membangun Tata Kelola Keamanan Produk Lintas Fungsi

Manajemen risiko tanggung jawab AI tidak boleh hanya berada di tangan satu tim saja, seperti tim teknik atau hukum. Sebaliknya, organisasi harus membentuk dewan peninjau produk multidisiplin sebagai bagian dari struktur tata kelola AI secara keseluruhan. Dewan peninjau tersebut akan melibatkan tim teknik, keamanan siber, privasi, manajemen risiko, hukum/kepatuhan, UX dan faktor manusia, serta spesialis risiko dan keselamatan. Dewan peninjau produk AI akan mengawasi keamanan produk, analisis dan mitigasi bahaya, pengungkapan peraturan, dan pengawasan pasca-pemasaran.

Sebelum meluncurkan pompa insulin berbasis AI, perusahaan perangkat medis melakukan Analisis Mode Kegagalan dan Dampak (FMEA), mengadakan rapat dewan keselamatan produk, dan menggunakan auditor pihak ketiga. Langkah-langkah ini mengurangi tanggung jawab desain dan pemasaran sekaligus selaras dengan harapan FDA.

Mengintegrasikan Pembaruan AI, Tata Kelola, dan Pemantauan Lapangan

Karena perilaku beberapa model AI dapat berubah setelah penerapan, tata kelola AI harus mencakup operasi berkelanjutan, termasuk pemantauan kinerja di dunia nyata, prosedur eskalasi kejadian dan insiden, dan rencana kontingensi yang mencakup pengembalian model dan pengesampingan manusia.

Organisasi harus mempertimbangkan untuk menerapkan peringatan otomatis untuk perubahan kinerja, pola yang tidak biasa, klaim kerugian, dan perilaku model yang tidak konsisten dengan hasil validasi. Regulator diharapkan meningkatkan harapan mereka untuk pengawasan pasca-pemasaran sistem berisiko tinggi, khususnya di bidang transportasi, perawatan kesehatan, dan infrastruktur.

Memodernisasi Pola Pikir Tanggung Jawab

Hukum tanggung jawab produk tradisional didasarkan pada produk fisik, desain statis, dan kausalitas linier. AI menantang semua asumsi ini, memperkenalkan otonomi, adaptivitas, dan inferensi, seringkali tanpa kendali langsung manusia.

Karakteristik ini tidak menghilangkan tanggung jawab; sebaliknya, sebagian besar justru mempersulitnya, karena peraturan dan yurisprudensi belum mampu mengimbangi dampak sosial AI.

Oleh karena itu, organisasi yang membangun atau menerapkan produk berbasis AI harus berasumsi bahwa kerugian akan terjadi dan menerapkan kontrol yang dapat dipertahankan sebelumnya. Kontrol ini dapat mencakup:

- Mendesain keamanan ke dalam persyaratan model dan sistem, arsitektur, desain, pengujian, dan operasi.
- Menguji di luar kondisi normal untuk mencakup kasus ekstrem, pengujian fuzzing, dan pergeseran.
- Memodelkan ancaman seluruh sistem AI ujung-ke-ujung.
- Mengkomunikasikan keterbatasan dan risiko dengan jelas kepada pengguna.
- Mendokumentasikan setiap langkah, dari persyaratan hingga desain hingga pembaruan.

Sistem hukum sedang beradaptasi tetapi lebih lambat daripada laju inovasi AI. Sementara itu, tim produk harus menjembatani kesenjangan antara kerangka hukum abad ke-20 dan sistem cerdas abad ke-21 dengan secara cermat mengantisipasi perkembangan dalam praktik, standar, peraturan, dan yurisprudensi.

5.5 RINGKASAN

Seiring semakin terintegrasinya sistem AI ke dalam sektor-sektor penting masyarakat, sistem ini bersinggungan dengan berbagai kerangka hukum yang ada di berbagai yurisdiksi dan industri. Bab ini mengkaji empat kategori utama hukum sektoral dan perdata (kekayaan intelektual, anti-diskriminasi, perlindungan konsumen, dan tanggung jawab produk) untuk memahami bagaimana masing-masing diterapkan pada desain, penerapan, dan penggunaan AI.

Hukum kekayaan intelektual memainkan peran mendasar dalam mengatur bagaimana sistem AI dilatih dan bagaimana outputnya dapat digunakan atau dilindungi. Hukum hak cipta menghadirkan tantangan baru, karena data pelatihan sering kali mencakup konten berhak cipta yang diambil dari web tanpa persetujuan pemiliknya. Meskipun pengembang berpendapat bahwa pelatihan merupakan penggunaan wajar, pengadilan semakin meneliti apakah penggunaan data hak cipta untuk melatih sistem AI melanggar hak yang ada. Hukum paten menimbulkan kekhawatiran terpisah tentang apakah penemuan yang dihasilkan AI dapat dipatenkan, dan siapa (atau apa) yang memenuhi syarat sebagai penemu.

Sementara itu, beberapa organisasi menggunakan perlindungan rahasia dagang untuk melindungi arsitektur model dan kumpulan data milik mereka. Secara global, perbedaan hak basis data dan pengecualian hak cipta mempersulit penerapan lintas batas. Organisasi harus melacak asal usul data, memperoleh lisensi yang tepat, dan mengelola risiko kekayaan intelektual melalui kontrol kontraktual dan teknis.

Hukum anti-diskriminasi mengatur penggunaan AI di berbagai bidang seperti pekerjaan, kredit, perumahan, dan asuransi. Undang-undang seperti Title VII dari Undang-

Undang Hak Sipil, Undang-Undang Kesempatan Kredit yang Setara, dan Undang-Undang Perumahan yang Adil melarang tidak hanya diskriminasi yang disengaja tetapi juga hasil yang tidak disengaja dengan dampak yang berbeda pada kelompok yang dilindungi.

Sistem AI yang bergantung pada data pelatihan yang bias atau menggunakan variabel proksi, seperti kode pos atau tingkat pendidikan, dapat secara tidak sengaja menghasilkan hasil yang diskriminatif, yang memicu risiko hukum. Regulator seperti EEOC dan CFPB telah mengklarifikasi bahwa tanggung jawab tetap berada pada organisasi, bahkan ketika alat AI dikembangkan oleh pihak ketiga (yang konsisten dengan doktrin manajemen risiko pihak ketiga yang telah ditetapkan). Hukum negara bagian dan lokal mencerminkan tren yang berkembang menuju deteksi dan tata kelola bias yang proaktif.

Undang-undang perlindungan konsumen, khususnya Bagian 5 dari Undang-Undang Komisi Perdagangan Federal (FTC), serta undang-undang UDAP negara bagian ("praktik tidak adil, menipu, dan menyalahgunakan"), melarang praktik tidak adil dan menipu dalam perdagangan. Undang-undang ini berlaku untuk sistem AI yang menyesatkan konsumen, mengeksploitasi kerentanan perilaku, atau mengaburkan fakta-fakta penting.

Risiko utama meliputi penggunaan AI yang disembunyikan, ulasan palsu, pola gelap, klaim yang dilebih-lebihkan, dan penipuan. FTC telah mengeluarkan panduan dan melakukan tindakan penegakan hukum terhadap perusahaan yang menyalahgunakan AI dalam pengenalan wajah, klaim kesehatan, dan penyimpanan data. Praktik terbaik meliputi memastikan transparansi, membuktikan klaim pemasaran terkait AI, dan menerapkan perlindungan bagi pengguna yang rentan secara emosional atau ekonomi.

Hukum tanggung jawab produk memperumit kompleksitas sistem AI yang berinteraksi dengan dunia fisik. Ketika kendaraan otonom, robot bedah, atau peralatan pintar menyebabkan kerugian, pengadilan dapat menerapkan teori yang ada tentang cacat desain, cacat manufaktur, atau kegagalan untuk memperingatkan. Namun, doktrin-doktrin ini kesulitan dengan perilaku adaptif, pengambilan keputusan probabilistik (vs. deterministik), dan rantai pasokan teknologi tinggi yang kompleks. Tanggung jawab dapat dibebankan kepada entitas mana pun yang terlibat dalam perancangan, integrasi, atau penerapan AI, tergantung pada sifat cacat dan kemungkinan terjadinya kerugian. Perusahaan harus secara proaktif menguji sistem AI dalam kondisi dunia nyata, memelihara dokumentasi terperinci, dan memperbarui pengungkapan risiko seiring perkembangan sistem.

Secara bersama-sama, rezim hukum ini menuntut pendekatan siklus hidup terhadap tata kelola AI yang mencakup penilaian risiko, pengawasan lintas fungsi, dan komitmen terhadap keadilan, keselamatan, dan transparansi di setiap tahap pengembangan dan pengoperasian AI.

BAB 6
UNDANG-UNDANG AI UNI EROPA

6.1 KLASIFIKASI RISIKO

Undang-Undang AI Uni Eropa memperkenalkan kerangka kerja terstruktur untuk mengklasifikasikan sistem AI berdasarkan risiko yang ditimbulkannya terhadap hak-hak fundamental, keselamatan, dan nilai-nilai masyarakat. Pendekatan berbasis risiko ini menekankan rasionalitas peraturan dan menentukan kewajiban yang harus dipatuhi oleh pengembang, penyedia, dan penyebar AI. Bagian ini membahas klasifikasi empat tingkat dalam Undang-Undang AI Uni Eropa: AI terlarang, AI berisiko tinggi, AI berisiko terbatas, dan AI berisiko minimal. Setiap tingkat klasifikasi memiliki persyaratan dan batasan yang sesuai. Memahami kategori-kategori ini sangat penting untuk penerapan AI yang sesuai dan etis di dalam Uni Eropa dan berpotensi di luar Uni Eropa.

Sistem AI yang Dilarang

Skema klasifikasi risiko Undang-Undang AI Uni Eropa mencakup satu kategori sistem AI yang tidak diizinkan.

Definisi dan Dasar Hukum

AI yang dilarang merujuk pada kasus penggunaan yang menimbulkan risiko yang tidak dapat diterima terhadap keselamatan, martabat manusia, atau hak-hak fundamental dan oleh karena itu dilarang sepenuhnya berdasarkan Undang-Undang AI Uni Eropa. Sistem ini tidak diizinkan untuk dipasarkan, dioperasikan, atau digunakan di mana pun di Uni Eropa.

Contoh AI yang Dilarang

Berikut adalah contoh sistem AI yang tidak diizinkan di Uni Eropa:

- ✓ Manipulasi subliminal: Sistem AI yang memanipulasi perilaku dengan cara yang melewati kesadaran, seperti teknologi yang memengaruhi neuro.
- ✓ Penilaian sosial oleh pemerintah: Sistem yang memberi peringkat individu berdasarkan perilaku atau ciri kepribadian, yang menyebabkan hasil yang diskriminatif.
- ✓ Mengeksploitasi kelompok rentan: Sistem AI yang menargetkan anak-anak, lansia, atau penyandang disabilitas untuk mendistorsi perilaku mereka atau orang lain.
- ✓ Identifikasi biometrik jarak jauh secara real-time di ruang publik: Dengan pengecualian terbatas, seperti untuk kejahatan serius dan di bawah pengawasan peradilan.
- ✓ Penilaian atau prediksi risiko tindak pidana individu.
- ✓ Pengakuan emosional di tempat kerja dan lembaga pendidikan: Secara tegas dilarang karena risiko manipulasi, diskriminasi, dan pseudosains.

Pengecualian terhadap Larangan Identifikasi Biometrik

Beberapa identifikasi biometrik secara real-time mungkin diizinkan dalam kondisi ketat, seperti ancaman keamanan nasional atau investigasi kriminal yang sedang berlangsung, tetapi harus diotorisasi secara individual oleh otoritas hukum.

Implikasi bagi Pengembang

Pengembang sistem AI (merujuk pada organisasi vendor, bukan insinyur) harus memastikan desain sistem, data pelatihan, dan kasus penggunaan mereka tidak melanggar wilayah terlarang. Pendekatan Uni Eropa mencerminkan sikap hati-hati terhadap penerapan AI yang dapat membahayakan individu atau masyarakat secara permanen.

Sistem AI Berisiko Tinggi

Skema klasifikasi Undang-Undang AI Uni Eropa mencakup tiga tingkatan sistem AI yang diizinkan. AI berisiko tinggi adalah risiko tertinggi dari ketiganya.

Definisi dan Kriteria

Sistem AI berisiko tinggi adalah sistem yang berpotensi secara signifikan memengaruhi hak, kehidupan, atau keselamatan masyarakat, dan karenanya tunduk pada kontrol peraturan yang ketat. Ini termasuk sistem yang digunakan di sektor yang diatur atau sistem yang memengaruhi keputusan penting.

Kategori AI Berisiko Tinggi (Lampiran III Undang-Undang):

- Identifikasi dan kategorisasi biometrik
- Manajemen infrastruktur kritis (misalnya, jaringan energi, pengelolaan air limbah, dan pengendalian lalu lintas)
- Pendidikan dan pelatihan kejuruan (misalnya, alat penilaian AI, perencanaan pelajaran)
- Ketenagakerjaan, manajemen pekerja, dan akses ke wirausaha (misalnya, wawancara AI, evaluasi kinerja berbasis AI)
- Akses dan pemanfaatan layanan swasta dan publik yang penting (misalnya, penilaian kredit, tunjangan kesejahteraan)
- Penegakan hukum (misalnya, alat prediksi kejahatan)
- Migrasi, suaka, dan kontrol perbatasan
- Administrasi peradilan dan proses demokrasi

Persyaratan Kepatuhan:

- **Penilaian kesesuaian:** Organisasi harus menunjukkan kepatuhan sebelum ditempatkan di pasar.
- **Manajemen risiko dan tata kelola data:** Pengembang harus menjaga kualitas data, meminimalkan bias, dan mengelola risiko operasional.
- **Pengawasan manusia:** Sistem harus memungkinkan intervensi dan eskalasi manusia yang berarti.
- **Dokumentasi teknis dan pencatatan:** Dokumentasi siklus hidup dan pencatatan otomatis harus dipelihara.

Penandaan CE untuk AI

Seperti penandaan keamanan produk lainnya, sistem AI berisiko tinggi harus memiliki penandaan CE untuk menunjukkan kesesuaian dengan standar Uni Eropa. Semua persyaratan akan berlaku.

Sebagai contoh, platform perekrutan tenaga kerja yang menggunakan AI untuk menyaring kandidat berdasarkan riwayat pekerjaan atau profil perilaku akan dianggap berisiko tinggi dalam kategori ketenagakerjaan.

Sistem AI Berisiko Terbatas

Skema klasifikasi Undang-Undang AI Uni Eropa mencakup tiga tingkatan sistem AI yang diizinkan. AI berisiko terbatas adalah tingkat risiko menengah dari ketiganya.

Definisi

AI berisiko terbatas merujuk pada sistem AI yang berinteraksi dengan pengguna tetapi tidak berdampak signifikan terhadap keselamatan atau hak.

Contoh AI berisiko terbatas meliputi:

- Chatbot dan asisten virtual
- Media yang dihasilkan AI (deepfake, gambar, atau suara sintetis)
- Alat pendidikan (asisten pembelajaran berbasis AI)
- Filter/avatar media sosial

Sistem-sistem ini memberlakukan kewajiban minimal yang terutama berfokus pada transparansi.

Kewajiban Transparansi

- **Pengungkapan penggunaan AI:** Pengguna harus diberi tahu ketika mereka berinteraksi dengan sistem AI.
- **Pengungkapan media sintetis:** Deepfake dan konten yang dihasilkan lainnya harus diberi label sebagai konten sintetis.
- **Otonomi pengguna:** Pengguna harus memiliki kesempatan untuk memilih keluar atau menolak interaksi AI tertentu jika wajar dan memungkinkan.

Misalnya, chatbot yang menyediakan dukungan pelanggan harus mengungkapkan bahwa ia bukan manusia. Demikian pula, gambar yang dihasilkan AI yang digunakan dalam iklan harus diberi label dengan jelas sebagai sintetis.

Pola Gelap dalam Sistem Berisiko Terbatas

Meskipun sistem ini membawa kewajiban terbatas, penggunaan pola gelap atau UI manipulatif dalam antarmuka AI dapat memicu pengawasan berdasarkan undang-undang perlindungan konsumen, yang terpisah dari Undang-Undang AI Uni Eropa dan GDPR. Pola gelap dibahas dalam Bab 4.

Sistem AI Berisiko Minimal

Skema klasifikasi Undang-Undang AI Uni Eropa mencakup tiga tingkatan sistem AI yang diizinkan. AI berisiko minimal adalah tingkat risiko terendah dari tiga tingkatan sistem AI yang diizinkan.

Definisi

Sistem AI berisiko minimal adalah sistem AI yang menimbulkan risiko yang dapat diabaikan atau tidak ada sama sekali bagi pengguna atau masyarakat. Sistem ini mewakili

sebagian besar sistem AI dan sebagian besar tidak diatur berdasarkan Undang-Undang AI Uni Eropa.

Contoh sistem AI berisiko minimal meliputi:

- Filter spam
- Mesin rekomendasi (misalnya, streaming video yang ditemukan di layanan streaming seperti Netflix, YouTube, dan Amazon Prime)
- Alat produktivitas (misalnya, pemeriksa tata bahasa yang didukung AI)

Praktik yang Dianjurkan

Meskipun tidak diwajibkan, pengembang sistem berisiko minimal didorong untuk mengikuti kode etik sukarela, pedoman etika, dan praktik transparansi, terutama ketika sistem mereka diskalakan atau diterapkan secara luas. Regulasi AI dapat berubah di masa mendatang dan mencakup pengawasan tambahan untuk sistem AI yang saat ini dianggap berisiko minimal.

Klasifikasi Ulang dan Pemantauan Risiko

Undang-Undang AI Uni Eropa mengantisipasi bahwa beberapa sistem AI dapat berevolusi dari waktu ke waktu atau digunakan kembali untuk tujuan lain. Sistem yang awalnya dikategorikan sebagai risiko minimal atau terbatas dapat menjadi risiko tinggi tergantung pada konteks penggunaannya. Pemantauan dan penilaian ulang secara terus-menerus disarankan.

Tabel 6.1 menyajikan perbandingan berdampingan dari skema klasifikasi risiko Undang-Undang AI Uni Eropa.

Tabel 6.1
Klasifikasi Risiko UU AI UE

Kategori Risiko	Definisi	Contoh	Persyaratan Peraturan
Dilarang AI	Menimbulkan risiko yang tidak dapat diterima terhadap hak/keselamatan	Penilaian sosial, manipulasi bawah sadar	Dilarang sepenuhnya
AI berisiko tinggi	Berdampak pada hak atau keamanan dasar	Penilaian kredit, AI alat perekrutan	Penilaian kesesuaian, pencatatan, pengawasan
AI dengan risiko terbatas	Membutuhkan transparansi tetapi dampaknya terbatas.	Chatbots, media sintetis	Pengungkapan, mekanisme penolakan
AI dengan risiko minimal	Menimbulkan risiko yang sangat kecil.	Pemeriksa ejaan, rekomendasi streaming	Tidak ada persyaratan (praktik terbaik sukarela)

Sistem klasifikasi ini merupakan inti dari pendekatan berbasis risiko dalam Undang-Undang AI Uni Eropa. Personel tata kelola AI harus memahami ambang batas dan batasan antara setiap tingkatan, yang sangat penting untuk memastikan kepatuhan hukum, mengelola risiko reputasi, dan membangun AI yang dapat dipercaya.

Persyaratan Klasifikasi

Kewajiban yang ditetapkan untuk setiap klasifikasi risiko AI berdasarkan Undang-Undang AI Uni Eropa bukan sekadar label; kewajiban tersebut mewakili harapan operasional

dan kepatuhan spesifik yang disesuaikan dengan profil risiko sistem. Bagian ini membahas persyaratan untuk sistem AI berisiko tinggi, berisiko terbatas, dan berisiko minimal, dengan penekanan khusus pada kontrol operasional seperti manajemen risiko, tata kelola data, dokumentasi, penilaian kesesuaian, langkah-langkah transparansi, dan pengawasan. Persyaratan ini mendefinisikan bagaimana sistem AI harus dirancang, divalidasi, diterapkan, dan dipantau.

Anda mungkin ingat bahwa pada bagian sebelumnya, empat kategori risiko telah dijelaskan. Namun, tidak ada persyaratan untuk klasifikasi AI terlarang, karena sistem ini sama sekali tidak diizinkan.

AI Berisiko Tinggi: Kerangka Kepatuhan Penuh

AI berisiko tinggi adalah klasifikasi risiko tertinggi untuk sistem AI dan, oleh karena itu, memiliki persyaratan yang paling ketat. Pembaca mungkin percaya bahwa keseluruhan persyaratan ini memberatkan. Namun, ini mewakili persyaratan serupa yang ditemukan dalam kerangka kerja siklus hidup pengembangan sistem yang matang. Dengan kata lain, jika Anda terkejut dengan banyaknya dokumen yang diperlukan untuk sistem AI berisiko tinggi, maka mungkin Anda belum memahami kerangka kerja pengembangan sistem yang matang.

Sistem Manajemen Risiko

Pengembang sistem AI berisiko tinggi harus menerapkan sistem manajemen risiko yang terdokumentasi dan berkelanjutan. Sistem ini harus:

- Secara proaktif mengidentifikasi risiko yang diketahui dan yang dapat diperkirakan terkait dengan penggunaan sistem AI yang dimaksudkan dan penyalahgunaan yang dapat diperkirakan secara wajar
- Mengevaluasi dan memantau risiko sepanjang siklus hidup sistem
- Menerapkan langkah-langkah pengendalian risiko untuk menghilangkan atau mengurangi risiko yang teridentifikasi

Beberapa kerangka kerja manajemen risiko yang baik tersedia untuk organisasi yang perlu menerapkan atau meningkatkan praktik manajemen risiko mereka, termasuk:

- ✓ NIST SP 800-37 (“Kerangka Kerja Manajemen Risiko untuk Sistem dan Organisasi Informasi”)
- ✓ ISO/IEC 27001 (“Keamanan Informasi, Keamanan Siber, dan Perlindungan Privasi—Sistem Manajemen Keamanan Informasi—Persyaratan”)
- ✓ ISO/IEC 31000 (“Manajemen Risiko—Pedoman”)
- ✓ ISO/IEC 23894:2023 (“Teknologi Informasi—Kecerdasan Buatan —Panduan tentang Manajemen Risiko”)

Perlu dicatat bahwa panduan hukum dalam standar NIST dan ISO ini bermanfaat tetapi bukan pengganti persyaratan eksplisit Undang-Undang AI Uni Eropa.

Data dan Tata Kelola Data

Sistem AI berisiko tinggi harus dilatih, divalidasi, dan diuji pada kumpulan data yang berkualitas tinggi, relevan, representatif, dan sebisa mungkin bebas dari kesalahan dan bias. Persyaratan meliputi:

- ➢ Strategi mitigasi bias

- Relevansi data dengan tujuan yang dimaksud
- Dokumentasi silsilah data dan kontrol versi

Organisasi yang melatih sistem AI dengan informasi pribadi juga harus mematuhi semua hukum privasi yang berlaku, termasuk GDPR Uni Eropa. Semua persyaratan terkait privasi harus diimplementasikan dalam sistem AI.

Organisasi yang tidak memiliki fungsi tata kelola data yang baik harus terlebih dahulu mengimplementasikan tata kelola data sebelum mengimplementasikan sistem AI berisiko tinggi. Organisasi yang membutuhkan panduan dapat merujuk pada sumber daya berikut:

- *Data Management Body of Knowledge* (DMBOK), yang mendefinisikan prinsip-prinsip tata kelola data, peran dan tanggung jawab, kebijakan dan prosedur, serta model kematangan tata kelola data.
- COBIT (sebelumnya, "*Control Objectives for Information and Related Technology*," sekarang hanya COBIT), yang membahas keselarasan bisnis, manajemen risiko, pengukuran kinerja, kepatuhan, dan persyaratan hukum.
- Kerangka Kerja *Data Governance Institute* (DGI), yang berisi empat komponen: (1) pernyataan nilai, tujuan, dan metrik; (2) aturan data, hak pengambilan keputusan, dan akuntabilitas; (3) kontrol dan standar; dan (4) orang dan proses.

Perusahaan konsultan, termasuk SAS, BCG, PWC, dan Eckerson, juga telah menerbitkan kerangka kerja tata kelola data.

Tata kelola data dibahas secara detail di Bab 9.

Dokumentasi Teknis

Penyedia sistem AI harus memelihara dokumentasi yang terperinci dan dikelola secara formal untuk menunjukkan kesesuaian dengan semua bagian yang berlaku dari Undang-Undang AI Uni Eropa, termasuk:

- Arsitektur dan spesifikasi sistem
- Algoritma dan metodologi pelatihan
- Protokol manajemen risiko dan pengawasan manusia

Untuk menghindari tumpang tindih dan duplikasi, organisasi harus mengembangkan dokumentasi teknis mereka untuk memenuhi semua persyaratan peraturan yang berlaku.

Pencatatan dan Log

Organisasi harus membuat dan memelihara log dan catatan lain sepanjang siklus hidup sistem AI. Log ini mendukung:

- Pemantauan pasca-pemasaran
- Investigasi peristiwa dan insiden
- Kesiapan dan akuntabilitas audit
- Manajemen perubahan pada tingkat proses bisnis dan sistem

Penilaian Kesesuaian

Sebelum dipasarkan, organisasi harus menjalani:

- Pengendalian internal (penilaian mandiri) dalam beberapa kasus
- Penilaian pihak ketiga oleh Badan yang Diberi Pemberitahuan dalam kasus lain
- Penandaan CE setelah kesesuaian dikonfirmasi

Dokumentasi dan catatan untuk kegiatan ini harus dicatat dan dipelihara.

Pengawasan Manusia

Sistem AI harus dirancang untuk memungkinkan:

- Deteksi anomali atau tanda-tanda kerusakan: Sebut ini perencanaan pemulihan bencana jika perlu: jika sistem AI (atau sistem informasi lainnya) mulai mengalami kerusakan, sistem tersebut harus dimatikan dan diperbaiki.
- Mekanisme penangguhan atau penghentian yang efektif: Jika situasi memungkinkan, organisasi harus dapat menghentikan sementara penggunaan sistem AI yang mengalami masalah, tanpa berdampak buruk pada proses bisnis yang penting.
- Tinjauan keputusan dengan campur tangan manusia jika diperlukan: Contohnya termasuk persetujuan pinjaman berbasis AI, penyaringan karyawan, diagnosis medis, dan banyak lagi.

Transparansi dan Instruksi Pengguna

Pengguna sistem AI harus menerima instruksi yang jelas untuk penggunaannya, termasuk:

- ✓ Tujuan dan batasan sistem
- ✓ Tanggung jawab pengawasan manusia
- ✓ Metrik kinerja dan kondisi penggunaan (tingkat akurasi, tingkat kesalahan yang diharapkan, persyaratan kualitas lingkungan atau data)
- ✓ Titik kontak untuk melaporkan insiden atau pertanyaan

Sistem Manajemen Mutu

Organisasi yang menerapkan sistem AI berisiko tinggi harus menerapkan sistem manajemen mutu (QMS), yang harus mencakup:

- ➢ Kebijakan dan prosedur untuk desain, pengembangan, pengujian, dan pengendalian
- ➢ Prosedur tindakan korektif dan preventif
- ➢ Pengendalian dokumentasi
- ➢ Proses untuk pemantauan, pelaporan, dan pencatatan selama operasi dan pengawasan pasca-pemasaran
- ➢ Manajemen sumber daya dan pelatihan untuk memastikan kompetensi

ISO 9000 (“Sistem Manajemen Mutu—Persyaratan”) adalah standar industri definitif untuk membangun sistem manajemen mutu.

Tanggung Jawab Dokumentasi Siklus Hidup

Pengembang sistem AI berisiko tinggi harus memelihara dokumentasi tidak hanya sebelum dipasarkan tetapi juga selama penggunaan sistem. Dokumentasi yang diperlukan mencakup pembaruan sistem, pelatihan ulang, dan adaptasi selanjutnya oleh pihak yang menerapkan sistem tersebut.

AI Risiko Terbatas: Kewajiban Transparansi yang Ditargetkan

AI risiko terbatas adalah klasifikasi risiko tingkat menengah untuk sistem AI dan, oleh karena itu, memiliki persyaratan yang kurang ketat daripada AI risiko tinggi.

Persyaratan Pengungkapan

Penyedia AI harus memastikan bahwa pengguna sistem AI berisiko terbatas diberi informasi:

- Saat berinteraksi dengan sistem AI (misalnya, chatbot dan asisten virtual)
- Saat melihat konten sintetis atau yang dimanipulasi (misalnya, deepfake, gambar, dan teks)

Pemberitahuan dan Otonomi Pengguna

Sistem AI berisiko terbatas harus mendukung:

- Pemberitahuan yang jelas bahwa sistem tersebut digerakkan oleh AI
- Cara yang mudah ditemukan bagi pengguna untuk memilih keluar atau beralih ke interaksi manusia jika sesuai
- Instruksi transparan tentang penggunaan dan batasan sistem sehingga pengguna memahami kemampuan dan batasannya

Dokumentasi dan Akuntabilitas

Meskipun tidak tunduk pada beban dokumentasi penuh dari sistem berisiko tinggi, pengembang AI berisiko terbatas harus:

- Memelihara catatan tentang bagaimana mekanisme transparansi diimplementasikan
- Memantau keluaran yang menyesatkan yang dapat mengganggu otonomi pengguna

Misalnya, situs ritel yang menggunakan chatbot penata gaya yang digerakkan oleh AI harus memberi tahu pengguna bahwa mereka berinteraksi dengan AI dan menawarkan opsi manusia jika memungkinkan.

Persimpangan Regulasi

Meskipun sistem berisiko terbatas diatur secara longgar berdasarkan Undang-Undang AI Uni Eropa, sistem tersebut mungkin masih berada di bawah hukum privasi dan/atau perlindungan konsumen Uni Eropa yang lebih luas, terutama ketika informasi pribadi atau pilihan desain yang menyesatkan/manipulatif terlibat.

AI Berisiko Minimal: Praktik Terbaik Sukarela

Tidak ada persyaratan wajib yang berlaku untuk sistem AI berisiko minimal. Namun, pengembang sistem AI berisiko minimal didorong untuk memberikan pemberitahuan sederhana di mana fungsi AI dapat memengaruhi keputusan pengguna, menghindari pola desain yang menyesatkan atau manipulatif, dan memelihara log internal dan pemeriksaan kualitas.

Sistem AI yang menimbulkan risiko minimal tetap harus mempertimbangkan akomodasi aksesibilitas untuk beragam pengguna dan menghindari penguatan stereotip atau disinformasi yang berbahaya. Misalnya, alat pengeditan tata bahasa yang ditingkatkan dengan AI dapat secara sukarela menyertakan pengungkapan bahwa saran tersebut dihasilkan oleh AI, sehingga meningkatkan kepercayaan tanpa tekanan regulasi.

Risiko Minimal Saat Ini, tetapi Bagaimana dengan Besok?

Sistem berisiko minimal yang diterapkan dalam skala besar atau diadaptasi ke konteks baru (misalnya, pendidikan atau kebijakan publik) dapat menghadapi klasifikasi ulang di masa mendatang. Praktik sukarela saat ini (terutama, mendokumentasikan detail desain, pengembangan, dan pelatihan sistem AI) dapat mengurangi biaya kepatuhan di kemudian hari.

Memahami persyaratan spesifik klasifikasi memungkinkan organisasi untuk menerapkan tata kelola yang proporsional, memastikan bahwa beban regulasi sebanding dengan potensi kerugian. Dengan demikian, hal ini meminimalkan regulasi berlebihan untuk alat-alat yang berdampak rendah sambil menerapkan perlindungan penuh pada sistem yang berisiko tinggi.

Hindari Kepatuhan Sekadar Centang Melalui Penyelarasan Bisnis

Banyak organisasi di setiap industri menunjukkan sikap minimalis terhadap kepatuhan, yang dikenal sebagai "kepatuhan centang," di mana kepemimpinan berupaya untuk mencapai hanya persyaratan minimum yang dinyatakan dengan cara yang paling hemat biaya. Seringkali, hal ini hanya menghasilkan penampilan kepatuhan yang tidak memberikan pengurangan risiko yang sebenarnya.

Pendekatan yang lebih sehat melibatkan mengejar semangat dan tujuan dari persyaratan yang dinyatakan, sambil juga mempertimbangkan langkah-langkah tambahan apa yang harus diterapkan organisasi untuk selaras dengan misi, tujuan, dan budayanya.

Persyaratan untuk AI Serbaguna

Seiring dengan pertumbuhan skala dan dampak sistem AI serbaguna (GPAI), Undang-Undang AI Uni Eropa memberlakukan serangkaian persyaratan khusus untuk mengatasi pengaruh sistemiknya. Tidak seperti alat AI yang ruang lingkupnya sempit, sistem GPAI dapat melayani berbagai aplikasi hilir, termasuk kasus penggunaan berisiko tinggi. Kompleksitas, fleksibilitas, dan ketersediaan umumnya menghadirkan tantangan tata kelola yang unik.

Bagian ini menguraikan persyaratan hukum, teknis, dan operasional spesifik yang berlaku untuk pengembang dan penyebar GPAI, dengan perbedaan berdasarkan apakah model tersebut bersifat sumber terbuka, komersial, atau ditetapkan memiliki risiko sistemik.

Definisi AI Serbaguna Berdasarkan Undang-Undang AI Uni Eropa

Definisi Hukum

Undang-Undang AI Uni Eropa mendefinisikan sistem AI serbaguna (GPAI) sebagai model yang dapat melayani berbagai tujuan dan diintegrasikan ke dalam berbagai sistem atau aplikasi hilir. Model-model ini tidak terikat pada satu tugas atau konteks penerapan tunggal. Contoh sistem GPAI meliputi:

- Model bahasa besar (LLM), seperti keluarga GPT, Llama, dan Claude
- Model dasar untuk tugas visi, seperti CLIP dan SAM
- Model multimodal yang mampu menginterpretasikan teks, gambar, video, dan audio

Model Dasar vs. AI Serbaguna

Meskipun sering digunakan secara bergantian, "model dasar" merujuk pada artefak teknis dasar, sedangkan "AI serbaguna" merujuk pada penggunaan dan ketersediaannya di berbagai aplikasi.

Kewajiban Umum bagi Penyedia GPAI

Penyedia GPAI diharuskan untuk mengembangkan dokumentasi, mengevaluasi dan menguji model AI mereka, serta mempraktikkan tata kelola data, sebagaimana dirinci dalam bagian ini.

Dokumentasi Teknis dan Transparansi

Semua penyedia GPAI harus menyiapkan dokumentasi teknis yang komprehensif, termasuk:

- Arsitektur model, sumber data pelatihan, dan kemampuan
- Penggunaan yang dimaksudkan dan yang dapat diperkirakan
- Keterbatasan, bias, dan mode kegagalan yang diketahui

Para penyedia ini juga harus mengungkapkan apakah model GPAI mereka telah disempurnakan pada data sensitif, serta mekanisme identifikasi dan mitigasi risiko, pemantauan, dan penonaktifan.

Evaluasi dan Pengujian Model

Penyedia GPAI harus melakukan dan mendokumentasikan penilaian risiko yang mengevaluasi:

- ✓ Akurasi dan ketahanan model AI di berbagai tugas
- ✓ Bias dan keadilan di berbagai demografi
- ✓ Dampak lingkungan (konsumsi energi)

Jika GPAI diharapkan untuk diintegrasikan ke dalam aplikasi berisiko tinggi, pengujian dan pengamanan tambahan diperlukan, sebagaimana diatur dalam skema klasifikasi risiko Undang-Undang AI Uni Eropa.

Tata Kelola Data

Meskipun penyedia GPAI tidak diwajibkan untuk sepenuhnya mengungkapkan dataset pelatihan mentah, penyedia harus mendokumentasikan komposisi dan asal usul dataset pelatihan, serta mengurangi risiko hak cipta, penyertaan data pribadi, dan bias representasional. Tata kelola data dibahas secara rinci di Bab 9.

Lembar Ringkasan Transparansi

Undang-Undang AI Uni Eropa mendorong pembuatan kartu model atau lembar fakta GPAI standar yang merangkum karakteristik dan risiko utama bagi pengguna hilir. Seperti yang dinyatakan di tempat lain dalam bab ini, organisasi harus mempertimbangkan untuk mendokumentasikan sepenuhnya pengembangan dan pelatihan sistem AI mereka, baik yang diwajibkan oleh peraturan maupun tidak.

Kewajiban Tambahan untuk GPAI dengan Risiko Sistemik

Menurut Undang-Undang AI Uni Eropa, model GPAI dengan risiko sistemik memiliki kewajiban kepatuhan tambahan.

Apa yang Termasuk dalam Risiko Sistemik?

Sistem GPAI dianggap menimbulkan risiko sistemik jika:

- Memiliki persyaratan komputasi yang sangat tinggi (misalnya, >1025 FLOPs; bukan batasan mutlak tetapi ilustrasi skala)
- Terintegrasi ke dalam infrastruktur kritis atau diterapkan secara luas di berbagai sektor industri

atau

- Telah menunjukkan kemampuan untuk secara otonom memengaruhi perilaku dalam skala besar

Persyaratan yang Ditingkatkan

Penyedia GPAI yang sistemnya memenuhi kriteria risiko sistemik harus:

- Menjalani audit eksternal
- Memberikan akses waktu nyata ke perilaku model—dalam kondisi tertentu
- Menerbitkan hasil pengujian yang merugikan
- Memberitahu regulator Uni Eropa tentang perubahan atau insiden signifikan
- Menerbitkan ringkasan transparansi tentang kemampuan, keterbatasan, dan risiko, sehingga pengguna hilir dan publik memahami dampak potensial

Tata Kelola dan Akuntabilitas

Penyedia GPAI dengan sistem AI risiko sistemik harus:

- Menetapkan penghubung hukum berbasis Uni Eropa
- Mempertahankan mekanisme respons dan eskalasi insiden
- Menyediakan API dan dokumentasi untuk peneliti keselamatan pihak ketiga

Perbandingan: GPAI Standar vs. GPAI Risiko Sistemik

Persyaratan	Standar GPAI	GPAI dengan Risiko Sistemik
Dokumentasi teknis	Ya	Ya
Audit eksternal	Tidak	Wajib
Pengujian permusuhan publikasi	Didorong	Diperlukan
Pemberitahuan regulator	Tidak	Diperlukan
Perwakilan hukum di UE	Tidak	Wajib

Kewajiban bagi Distributor dan Penyebar GPAI

Organisasi yang mengintegrasikan GPAI ke dalam sistem hilir, terutama aplikasi berisiko tinggi, diharuskan untuk:

- Memverifikasi kesesuaian model dengan dokumentasi yang dipublikasikan
- Menilai kasus penggunaan mereka sendiri untuk potensi peningkatan klasifikasi risiko
- Memastikan transparansi terhadap pengguna akhir

Sebagai contoh, bank yang berbasis di Uni Eropa yang mengintegrasikan model GPAI ke dalam platform penjaminan pinjamannya harus memverifikasi:

- Keakuratan dan kelengkapan dokumentasi model dan profil risikonya
- Bahwa pengawasan manusia tetap dipertahankan
- Bahwa bank tersebut mematuhi semua kewajiban sistem berisiko tinggi jika dipicu

Pertimbangan GPAI Sumber Terbuka

Pengembang GPAI sumber terbuka mungkin dibebaskan dari beberapa kewajiban jika mereka tidak memperoleh manfaat komersial atau mempertahankan kendali operasional atas sistem yang mereka kembangkan. Namun, pihak yang menerapkan tetap bertanggung jawab atas kepatuhan di hilir.

Praktik yang Dianjurkan di Semua Model GPAI

Meskipun kewajiban tidak bersifat wajib, Undang-Undang AI Uni Eropa mendorong adopsi kode etik sukarela yang membahas aksesibilitas, inklusivitas, keberlanjutan lingkungan, dan pencegahan penyalahgunaan (misalnya, manipulasi pemilihan, pembuatan malware, atau penipuan). Lebih lanjut, Undang-Undang tersebut mendorong pengembang GPAI untuk berkolaborasi dengan akademisi dan masyarakat sipil serta mendukung tolok ukur terbuka dan inisiatif transparansi.

Fleksibilitas GPAI memerlukan pendekatan tata kelola yang berbeda dibandingkan dengan alat AI fungsi tetap, karena sistem GPAI dapat diadaptasi ke berbagai lingkungan dan mendukung beragam kasus penggunaan, termasuk beberapa yang mungkin membawa risiko tinggi. Dengan mengatur GPAI melalui ketentuan umum dan yang dipicu risiko, Undang-Undang AI Uni Eropa bertujuan untuk memastikan sistem yang ampuh ini dikembangkan dan diterapkan secara bertanggung jawab, bahkan ketika aplikasinya belum sepenuhnya diketahui.

6.2 PENEGAKAN DAN SANKSI

Peraturan yang tidak memiliki konsekuensi atas ketidakpatuhan memiliki sedikit nilai. Undang-Undang AI Uni Eropa menetapkan penegakan yang kuat untuk memastikan kepatuhan terhadap ketentuannya di semua tingkatan risiko, konteks, dan peran organisasi. Penegakan hukum dirancang agar proporsional dengan tingkat risiko yang ditimbulkan oleh suatu sistem dan tingkat keparahan ketidakpatuhan. Bagian ini menjelaskan bagaimana peraturan akan dipantau dan ditegakkan oleh otoritas Uni Eropa dan nasional, dan menjelaskan sanksi yang mungkin dihadapi organisasi jika mereka gagal mematuhi.

Penegakan hukum berdasarkan Undang-Undang AI Uni Eropa mencerminkan keseimbangan antara inovasi dan akuntabilitas. Dengan menetapkan kewenangan pengawasan yang kuat dan sanksi yang bersifat pencegahan, Undang-Undang ini memperkuat pendekatan berbasis risiko, dengan meminta pertanggungjawaban yang lebih besar dari pengembang dan penyedia di mana penggunaan AI menimbulkan bahaya yang lebih besar bagi hak, keselamatan, atau demokrasi.

Pengawasan dan Otoritas Pengawasan

Undang-Undang AI Uni Eropa menyediakan dua tingkat otoritas penegakan hukum—Kantor AI Eropa dan negara-negara anggota.

Undang-Undang AI Uni Eropa menetapkan badan terpusat yang dikenal sebagai Kantor AI Eropa, yang berada di bawah Komisi Eropa. Badan ini diharuskan untuk mengkoordinasikan pengawasan sistem dan model GPAI dengan risiko sistemik, memelihara basis data publik sistem AI berisiko tinggi, dan mengeluarkan panduan dan praktik terbaik di seluruh negara anggota.

Setiap negara anggota diwajibkan untuk menunjuk satu atau lebih otoritas kompeten nasional untuk melakukan pengawasan pasar, menangani pengaduan dari individu dan masyarakat sipil, menyelidiki pelanggaran, dan menjatuhkan sanksi dalam yurisdiksi negara anggota tersebut.

Koordinasi Antara Uni Eropa dan Negara-Negara Anggota
Meskipun penegakan hukum didorong secara nasional, Kantor AI Eropa memainkan peran harmonisasi dan pengawasan sentral, khususnya untuk GPAI dan penerapan AI lintas batas.

Kewenangan dan Upaya Inspeksi

Sebagai bagian dari pengawasan pasar, otoritas dapat meminta akses ke dokumentasi teknis, melakukan inspeksi mendadak (termasuk audit digital), mewajibkan dan memverifikasi tindakan korektif, dan menangguhkan atau melarang pemasaran sistem AI.

Tergantung pada tingkat keparahan hasil audit atau pelanggaran, otoritas dapat meminta agar sistem yang tidak sesuai ditarik atau ditarik kembali. Mereka juga dapat mengarahkan pengembang atau penyebar AI untuk melakukan pembaruan, melatih ulang sistem mereka, atau memberi label pada perubahan sistem. Mereka dapat meminta penyedia untuk memberi tahu pengguna yang terpengaruh.

Individu yang terdampak oleh sistem AI yang tidak sesuai standar mungkin dapat mencari ganti rugi berdasarkan kerangka tanggung jawab perdata nasional, atau mengajukan pengaduan melalui saluran perlindungan data atau perlindungan konsumen.

Perlindungan Pelapor Pelanggaran
Undang-Undang AI Uni Eropa mendorong pelaporan ketidakpatuhan sistemik dan dapat mencakup ketentuan untuk melindungi pelapor pelanggaran di dalam organisasi AI, serupa dengan ketentuan dalam GDPR dan undang-undang Uni Eropa lainnya.

Sanksi untuk Ketidakpatuhan

Undang-Undang AI Uni Eropa memperkenalkan kerangka sanksi bertingkat berupa denda administratif sebagai berikut:

- Hingga Rp 350 miliar atau 7% dari omset tahunan global (mana yang lebih tinggi) untuk pelanggaran praktik terlarang (misalnya, AI yang dilarang)
- Hingga Rp 150 miliar atau 3% dari omset (mana yang lebih tinggi) untuk ketidakpatuhan terhadap kewajiban untuk sistem berisiko tinggi

- Hingga Rp 75 miliar juta atau 1% dari omset (mana yang lebih tinggi) untuk memberikan informasi yang salah, tidak lengkap, atau menyesatkan kepada pihak berwenang

Sanksi Dibandingkan dengan GDPR
Sanksi dalam Undang-Undang AI Uni Eropa menyerupai sanksi berdasarkan GDPR tetapi dapat melebihi sanksi tersebut untuk pelanggaran paling berat yang melibatkan AI terlarang atau risiko sistemik GPAI.

Misalnya, jika penyedia alat perekrutan berbasis AI gagal menerapkan mitigasi bias atau pengawasan manusia yang diperlukan, dan ini menyebabkan hasil yang diskriminatif, mereka dapat menghadapi denda hingga 315 juta, ditambah penarikan wajib sistem tersebut.

Eskalasi dan Pengulangan

Undang-Undang AI Uni Eropa mencakup ketentuan untuk masalah kepatuhan yang tidak mudah diselesaikan.

Dalam menilai sanksi, pihak berwenang akan mempertimbangkan:

- ✓ Sifat dan signifikansi pelanggaran
- ✓ Apakah ada masalah ketidakpatuhan, dan, jika demikian, apakah disengaja atau lalai
- ✓ Riwayat pelanggaran sebelumnya
- ✓ Tingkat kerja sama dengan pihak berwenang

Setelah tindakan penegakan hukum, organisasi yang dikenai sanksi mungkin juga diharuskan untuk menjalani audit pihak ketiga, menerapkan rencana perbaikan kepatuhan, dan melaporkan kemajuan perbaikan secara berkala.

Manfaat Pengungkapan Sukarela

Organisasi yang secara sukarela mengungkapkan insiden atau kekurangan dan bekerja sama dengan pihak berwenang dapat menerima pengurangan hukuman atau menghindari sanksi yang bersifat menghukum.

Penegakan dan Kerja Sama Lintas Batas

Undang-Undang AI Uni Eropa mencakup ketentuan terkait masalah penegakan hukum yang melibatkan dua negara anggota atau lebih, serta pengembang dan penyedia di luar Uni Eropa.

Kerangka Kerja Saling Membantu

Undang-Undang AI Uni Eropa mengarahkan otoritas nasional untuk saling membantu ketika masalah penegakan hukum melintasi batas negara. Otoritas nasional akan bekerja sama melalui:

- ➢ Berbagi informasi tentang investigasi
- ➢ Mendukung tindakan penegakan hukum lintas batas
- ➢ Mengkoordinasikan kegiatan ini dan kegiatan lainnya melalui Kantor AI Eropa

Kejelasan Yurisdiksi untuk Penyedia GPAI

Pengembang GPAI di luar Uni Eropa diharuskan untuk menunjuk perwakilan hukum yang berbasis di Uni Eropa untuk mewakili organisasi tersebut, di mana pun kantor pusatnya berada. Pengembang GPAI juga diharuskan untuk memastikan semua sistem yang ditempatkan di pasar Uni Eropa memenuhi semua standar kepatuhan yang berlaku.

Ekstrateritorialitas

Undang-Undang AI Uni Eropa berlaku tidak hanya untuk organisasi yang berbasis di Uni Eropa tetapi juga untuk entitas non-Uni Eropa jika mereka menempatkan sistem AI di pasar Uni Eropa atau menyediakan sistem AI kepada pengguna yang berlokasi di Uni Eropa. Ini termasuk sistem AI dan sistem AI tujuan umum yang dipasarkan atau digunakan di Uni Eropa, dan pihak yang menerapkan di luar Uni Eropa jika sistem tersebut memengaruhi warga negara Uni Eropa. Regulasi ini juga memengaruhi pengembang GPAI di luar Uni Eropa ketika model mereka diintegrasikan ke dalam aplikasi yang ditujukan untuk Uni Eropa (bahkan jika aplikasi tersebut tidak berada secara fisik di Uni Eropa).

Penyedia non-Uni Eropa diharuskan untuk menunjuk perwakilan hukum yang berbasis di Uni Eropa, memastikan semua penilaian kesesuaian telah dilakukan, dan memenuhi semua persyaratan transparansi dan dokumentasi berdasarkan klasifikasi sistem.

6.3 KONTEKS ORGANISASI

Kepatuhan terhadap Undang-Undang AI Uni Eropa tidak hanya bergantung pada jenis sistem AI yang bersangkutan, tetapi juga pada peran yang dimainkan suatu organisasi dalam rantai pasokan AI. Apakah suatu entitas merupakan penyedia, pihak yang menerapkan, pengimpor, atau distributor secara signifikan membentuk kewajiban hukumnya. Bagian ini menjelaskan persyaratan yang berbeda untuk setiap peran, menyoroti bagaimana konteks organisasi mendorong akuntabilitas berdasarkan regulasi tersebut.

Penyedia AI: Beban Kepatuhan Inti

Penyedia adalah setiap orang atau badan hukum yang mengembangkan sistem AI atau model GPAI dengan maksud untuk memasarkannya di pasar Uni Eropa, dengan nama atau merek sendiri.

Kewajiban utama penyedia meliputi:

- Melakukan dan mendokumentasikan prosedur manajemen risiko
- Menyelesaikan dan mendokumentasikan penilaian kesesuaian
- Memelihara dokumentasi dan log teknis
- Menerapkan penandaan CE (untuk sistem AI berisiko tinggi)
- Mendaftarkan sistem di basis data AI Uni Eropa
- Memastikan pemantauan pasca-pemasaran dan pelaporan insiden

Penyedia GPAI vs. AI Sempit

Penyedia GPAI menghadapi kewajiban tambahan dan berbeda, termasuk transparansi model, analisis risiko sistemik, dan dokumentasi kemampuan dan keterbatasan, di luar apa yang harus dilakukan oleh penyedia AI sempit tradisional.

Contoh perusahaan yang melatih dan menjual model pengenalan wajah untuk penggunaan komersial adalah penyedia dan memikul tanggung jawab penuh atas dokumentasi, pengendalian risiko, dan transparansi.

Pengguna: Akuntabilitas Operasional dalam Praktik

Pengguna adalah orang atau organisasi yang menggunakan sistem AI dalam konteks aktivitas profesional mereka. Ini termasuk organisasi publik dan swasta yang menerapkan atau memasang sistem AI yang dikembangkan oleh pihak lain. Anggaplah pengguna sebagai orang, tim, atau organisasi yang menerapkan sistem AI untuk aktivitas profesionalnya (bukan untuk penggunaan pribadi).

Kewajiban pengguna bervariasi tergantung pada jenis sistem AI.

AI berisiko tinggi:

- Pastikan pengawasan manusia diterapkan selama penggunaan sistem.
- Memantau kinerja sistem dan melaporkan insiden serius
- Memastikan kualitas data masukan dan kesesuaiannya dengan konteks lokal
- Memelihara log sistem dan catatan penggunaan
- Menggunakan sistem sesuai dengan instruksi penyedia

Integrasi GPAI:

- Validasi dokumentasi dan uji kompatibilitas
- Pertimbangkan peran dalam reklasifikasi risiko jika mengadaptasi atau menggunakan kembali AI

Ketika Pihak yang Menerapkan Menjadi Penyedia

Jika pihak yang menerapkan secara substansial memodifikasi sistem AI atau mengembangkan aplikasi hilir sendiri menggunakan GPAI, secara hukum pihak tersebut dapat diklasifikasi ulang sebagai penyedia dan mewarisi kewajiban tersebut.

Contoh pihak yang menerapkan sistem adalah rumah sakit yang menerapkan alat diagnostik berbasis AI, yang harus memastikan bahwa protokol pengawasan telah diterapkan dan sistem tersebut beroperasi dengan aman di lingkungan klinis.

Importir: Menjembatani Yurisdiksi Hukum

Importir adalah organisasi yang menempatkan sistem AI dari negara-negara non-UE ke pasar UE. Mereka bertindak sebagai perantara hukum antara penyedia asing dan regulator yang berbasis di UE.

Kewajiban importir meliputi:

- Memverifikasi tanda CE dan penilaian kesesuaian sistem
- Memastikan dokumentasi teknis tersedia, benar, dan lengkap

- Mendaftarkan sistem AI di basis data UE (kecuali dilakukan oleh penyedia)
- Bekerja sama dengan otoritas terkait pemeriksaan kepatuhan

Paparan Risiko Importir

Importir bertanggung jawab secara hukum jika mereka memasarkan sistem AI yang tidak sesuai standar. Kontrak antara importir dan penyedia harus memperjelas batasan tanggung jawab.

Contohnya, pengecer teknologi yang berbasis di Jerman mengimpor alat AI pendidikan yang dikembangkan di AS ke pasar Uni Eropa. Pengecer tersebut, dalam peran sebagai importir, harus memastikan bahwa semua dokumentasi, registrasi, dan tanda peraturan telah tersedia.

Distributor: Pasif tetapi Tidak Dikecualikan

Distributor menyediakan sistem AI di pasar Uni Eropa tanpa memodifikasinya atau bertindak sebagai importir. Contohnya termasuk toko aplikasi, pengecer platform, dan pasar online.

Kewajiban distributor berdasarkan Undang-Undang AI Uni Eropa meliputi:

- Memverifikasi keberadaan tanda CE
- Mengkonfirmasi informasi kontak penyedia/importir terlihat
- Menangguhkan distribusi jika dicurigai adanya ketidakpatuhan
- Bekerja sama dengan pihak berwenang selama investigasi

Contoh platform e-commerce yang mencantumkan perangkat rumah bertenaga AI adalah distributor menurut Undang-Undang AI Uni Eropa. Organisasi tersebut harus menghapus daftar produk jika menerima bukti ketidaksesuaian yang kredibel.

Distributor, Pemberi Pemberitahuan, dan Pelaporan Pelanggaran

Distributor didorong (tetapi tidak diwajibkan) untuk menetapkan mekanisme bagi konsumen atau pesaing untuk melaporkan potensi ketidakpatuhan atau risiko AI.

Tanggung Jawab Bersama dan Tumpang Tindih

Mari kita lihat sekilas situasi tanggung jawab bersama dan tumpang tindih.

Tugas Horizontal

Di semua peran, organisasi diharuskan untuk:

- ✓ Bekerja sama dengan otoritas dan regulator nasional yang berwenang
- ✓ Menanggapi laporan insiden
- ✓ Memelihara catatan (log, pembaruan, pemberitahuan)

AI sebagai Layanan dan Ambiguitas Peran

Platform AI berbasis cloud dapat mengaburkan peran:

- ➢ Apakah penyedia cloud adalah penyebar, pengimpor, distributor, atau pengguna
- ➢ Pihak mana yang melakukan pemeriksaan kesesuaian pada layanan AI modular

Kejelasan mungkin memerlukan penetapan tanggung jawab kontraktual yang terkait dengan peran yang didefinisikan dalam Undang-Undang AI Uni Eropa.

Memetakan Peran Anda dalam Siklus Hidup AI

Organisasi harus melakukan "penilaian peran AI" untuk menentukan apakah mereka bertindak sebagai penyedia, penyebar, pengimpor, atau distributor untuk setiap sistem—dan kewajiban kepatuhan apa yang dihasilkan.

Memahami beban regulasi yang terkait dengan setiap peran organisasi memastikan tidak hanya kepatuhan tetapi juga akuntabilitas yang lebih jelas, kesiapan tata kelola, dan koordinasi rantai pasokan. Undang-Undang AI Uni Eropa menuntut agar semua pihak, dari pengembang hingga pengecer, menerima dan mendokumentasikan tanggung jawab masing-masing untuk AI yang dapat dipercaya.

6.4 IMPLEMENTASI UNDANG-UNDANG AI UNI EROPA

Undang-Undang AI Uni Eropa mengikuti jadwal implementasi bertahap, dengan kewajiban dan masa tenggang yang berbeda-beda, tergantung pada jenis sistem AI dan peran pemangku kepentingan. Tabel 6.2 berisi uraian singkat.

Tabel 6.2

Tonggak Penting dalam Undang-Undang AI Uni Eropa

Tanggal	Acara/Kewajiban
12 Juli 2024	Undang-Undang AI diterbitkan dalam Jurnal Resmi Uni Eropa.
1 Agustus 2024	Berlaku efektif - dimulainya pemberlakuan peraturan secara resmi tetapi tidak ada kewajiban langsung.
2 Februari 2025	Praktik AI yang dilarang dan persyaratan literasi AI mulai berlaku.
2 Mei 2025	Pedoman praktik untuk model GPAI diharapkan akan segera diselesaikan.
2 Agustus 2025	Kewajiban penting dimulai: aturan GPAI, struktur tata kelola, sanksi, badan yang diberi wewenang.
2 Agustus 2025	Penyedia model GPAI yang sudah ada di pasaran harus mematuhi ketentuan ini paling lambat tanggal 2 Agustus 2027.
2 Agustus 2026	Persyaratan sistem AI berisiko tinggi mulai berlaku (misalnya, penilaian kesesuaian, penandaan CE).
Pada tahun 2027	Diharapkan penegakan penuh terhadap semua ketentuan yang berlaku.

Pembaca disarankan untuk secara berkala meninjau berita mengenai penegakan Undang-Undang AI Uni Eropa, karena tanggal-tanggal ini dapat berubah.

Berikut beberapa masa tenggang dan ketentuan transisi:

- **Penyedia GPAI:** Jika model GPAI dipasarkan sebelum 2 Agustus 2025, penyedia memiliki waktu hingga 2 Agustus 2027 untuk mencapai kepatuhan penuh.
- **Kode praktik:** Meskipun bersifat sukarela, menandatangani Kode Praktik GPAI dapat memberikan jaminan hukum. Pihak yang tidak menandatangani tidak akan mendapatkan manfaat dari asumsi kesesuaian.

- **Negara-negara anggota:** Mereka harus menunjuk otoritas penegak hukum dan menetapkan kerangka sanksi paling lambat tanggal 2 Agustus 2025.

Meta dan Kode Etik GPAI

Pada Juli 2025, Meta mengumumkan bahwa perusahaan tersebut tidak akan menandatangani Kode Etik GPAI, dengan alasan ketidakpastian hukum, campur tangan regulasi yang berlebihan, dan kekhawatiran atas inovasi AI. Penyedia AI lainnya, termasuk OpenAI dan Mistral AI, telah setuju untuk menandatangani Kode Etik tersebut.
Hingga saat penulisan buku ini, masalah ini terus berkembang.

6.5 RINGKASAN

UU AI Uni Eropa memperkenalkan kerangka kerja berbasis risiko bertingkat untuk mengatur pengembangan dan penggunaan sistem AI, mengklasifikasikannya ke dalam empat kategori: dilarang, berisiko tinggi, berisiko terbatas, dan berisiko minimal. Setiap kategori sesuai dengan tingkat bahaya yang dapat ditimbulkan sistem tersebut terhadap hak-hak fundamental, keselamatan, atau nilai-nilai demokrasi. Sistem AI yang dilarang mencakup praktik-praktik seperti penilaian sosial oleh pemerintah atau manipulasi subliminal, dan dilarang sepenuhnya.

Sistem AI berisiko tinggi (yang memengaruhi area kritis seperti pekerjaan, pendidikan, atau penegakan hukum) tunduk pada kontrol yang paling ketat, termasuk penilaian kesesuaian, dokumentasi, pengawasan, dan penandaan CE. Sistem AI berisiko terbatas, seperti chatbot dan deepfake, harus memenuhi persyaratan transparansi. Sistem AI berisiko minimal seperti filter spam dan pemeriksa tata bahasa sebagian besar tidak diatur, meskipun mereka didorong untuk mendokumentasikan pengembangan, manajemen, dan proses operasionalnya.

Setiap tingkatan risiko memiliki persyaratan peraturan tersendiri. AI berisiko tinggi harus menerapkan protokol manajemen risiko yang kuat, memastikan data pelatihan berkualitas tinggi, memelihara dokumentasi teknis yang komprehensif, dan mendukung pengawasan manusia yang bermakna. Sistem ini juga membutuhkan sistem manajemen mutu dan pemantauan pasca-pemasaran. Sistem AI berisiko terbatas terutama terikat oleh kewajiban transparansi dan pemberitahuan untuk memberi tahu pengguna bahwa mereka berinteraksi dengan AI. Untuk sistem AI berisiko minimal, Undang-Undang tersebut mendorong (tetapi tidak mewajibkan) praktik baik sukarela seputar transparansi dan etika. Di semua tingkatan, peraturan tersebut mendorong proporsionalitas dan pemikiran siklus hidup.

Sistem AI tujuan umum (GPAI), seperti model bahasa besar, menerima perlakuan regulasi khusus dan tambahan berdasarkan Undang-Undang AI Uni Eropa. Penyedia GPAI harus mengungkapkan arsitektur model, praktik data pelatihan, keterbatasan yang diketahui, dan tujuan penggunaan. Mereka juga diharapkan untuk melakukan pengujian ketahanan, keadilan, dan dampak lingkungan.

Model yang dianggap menimbulkan risiko sistemik (berdasarkan skala, dampak, atau pengaruh sosial) harus memenuhi standar yang lebih tinggi, termasuk audit eksternal,

publikasi pengujian yang merugikan, dan pemberitahuan kepada regulator. Undang-Undang tersebut juga membebankan kewajiban pada organisasi yang menerapkan atau mengintegrasikan GPAI ke dalam sistem hilir, terutama ketika penggunaan tersebut menghasilkan penetapan risiko tinggi atau risiko terbatas.

Undang-Undang AI membedakan kewajiban berdasarkan peran organisasi suatu entitas, apakah itu penyedia, penyebar, importir, atau distributor. Penyedia memikul tanggung jawab terbesar, termasuk melakukan penilaian risiko, menyelesaikan pemeriksaan kesesuaian, dan memastikan kepatuhan pasca-pemasaran. Penyebar harus menerapkan pengawasan, mengelola kinerja sistem, dan menanggapi hasil yang merugikan. Importir memastikan sistem asing memenuhi kepatuhan UE sebelum memasuki pasar, dan distributor harus bertindak jika mereka mengetahui adanya ketidakpatuhan. Perbedaan ini memperjelas akuntabilitas di seluruh siklus hidup AI dan membantu memastikan integritas rantai pasokan. Suatu organisasi dapat memiliki lebih dari satu peran; misalnya, suatu organisasi dapat menjadi importir dan distributor sekaligus.

Undang-Undang AI UE mencakup kerangka kerja penegakan hukum yang komprehensif. Kantor AI Eropa memimpin pengawasan terpusat, terutama untuk GPAI, sementara otoritas nasional menangani inspeksi, pengawasan pasar, dan sanksi.

Denda untuk ketidakpatuhan bertingkat berdasarkan sifat dan tingkat keparahan pelanggaran, mulai dari hingga 335 juta atau 7% dari pendapatan global (mana yang lebih besar) untuk praktik terlarang. Otoritas dapat menangguhkan atau menarik kembali sistem, meminta perubahan, atau mewajibkan pemberitahuan kepada pengguna. Entitas yang terbukti melanggar juga dapat menghadapi pemantauan berkelanjutan atau audit pihak ketiga. Undang-Undang ini mendorong pengungkapan dan kerja sama sukarela, menawarkan potensi pengurangan sanksi sebagai insentif.

Organisasi harus berasumsi bahwa Undang-Undang AI Uni Eropa, serta peraturan AI lain yang berlaku, akan berubah seiring waktu, sehingga memberlakukan persyaratan baru atau berbeda pada organisasi dalam berbagai peran rantai pasokan AI mereka.

BAB 7
STANDAR DAN KERANGKA KERJA AI

7.1 PRINSIP-PRINSIP AI OECD

Prinsip-Prinsip AI OECD, yang diadopsi pada tahun 2019 oleh negara-negara anggota OECD (Organisasi untuk Kerja Sama dan Pembangunan Ekonomi), adalah serangkaian pedoman awal untuk AI yang dapat dipercaya. Prinsip-prinsip ini memandu desain dan penerapan sistem AI yang bertanggung jawab. Meskipun tidak mengikat secara hukum, prinsip-prinsip ini telah memengaruhi banyak strategi AI nasional dan internasional serta pendekatan regulasi, termasuk Undang-Undang AI Uni Eropa dan Proses Hiroshima G7.

Anda dapat menemukan informasi lebih lanjut tentang Prinsip-Prinsip AI OECD di *https://oecd.ai/en/ai-principles*.

Prinsip-Prinsip Utama OECD untuk Pengelolaan AI yang Terpercaya dan Bertanggung Jawab

Berikut adalah lima prinsip AI yang terpercaya:

- **Pertumbuhan inklusif, pembangunan berkelanjutan, dan kesejahteraan:** Mengutip OECD, "Para pemangku kepentingan harus secara proaktif terlibat dalam pengelolaan AI yang terpercaya dan bertanggung jawab untuk mencapai hasil yang bermanfaat bagi manusia dan planet ini, seperti meningkatkan kemampuan manusia dan meningkatkan kreativitas, memajukan inklusi populasi yang kurang terwakili, mengurangi ketidaksetaraan ekonomi, sosial, gender, dan lainnya, serta melindungi lingkungan alam, sehingga mendorong pertumbuhan inklusif, kesejahteraan, pembangunan berkelanjutan, dan keberlanjutan lingkungan."
- **Hak asasi manusia dan nilai-nilai demokrasi, termasuk keadilan dan privasi:** "Para pelaku AI harus menghormati supremasi hukum, hak asasi manusia, nilai-nilai demokrasi dan berpusat pada manusia di seluruh siklus hidup sistem AI. Ini termasuk non-diskriminasi dan kesetaraan, kebebasan, martabat, otonomi individu, privasi dan perlindungan data, keragaman, keadilan, keadilan sosial, dan hak-hak buruh yang diakui secara internasional. Ini juga termasuk mengatasi misinformasi dan disinformasi yang diperkuat oleh AI, sambil menghormati kebebasan berekspresi dan hak serta kebebasan lain yang dilindungi oleh hukum internasional yang berlaku."
- **Transparansi dan penjelasan:** "Para pelaku AI harus berkomitmen pada transparansi dan pengungkapan yang bertanggung jawab terkait sistem AI. Untuk tujuan ini, mereka harus memberikan informasi yang bermakna, sesuai dengan konteks, dan konsisten dengan perkembangan terkini:
  - Untuk menumbuhkan pemahaman umum tentang sistem AI, termasuk kemampuan dan keterbatasannya.
  - Untuk membuat para pemangku kepentingan menyadari interaksi mereka dengan sistem AI, termasuk di tempat kerja.

  - Jika memungkinkan dan bermanfaat, untuk memberikan informasi yang jelas dan mudah dipahami tentang sumber data/masukan, faktor, proses dan/atau logika yang mengarah pada prediksi, konten, rekomendasi atau keputusan, untuk memungkinkan mereka yang terpengaruh oleh sistem AI memahami hasilnya.
  - Untuk memberikan informasi yang memungkinkan mereka yang terkena dampak negatif oleh sistem AI untuk menantang hasilnya."
- **Ketangguhan, keamanan, dan keselamatan:**
  - "Sistem AI harus tangguh, aman, dan terlindungi sepanjang siklus hidupnya sehingga, dalam kondisi penggunaan normal, penggunaan yang dapat diprediksi, atau penyalahgunaan, atau kondisi buruk lainnya, sistem tersebut berfungsi dengan baik dan tidak menimbulkan risiko keselamatan dan/atau keamanan yang tidak wajar.
  - Mekanisme harus tersedia, sebagaimana mestinya, untuk memastikan bahwa jika sistem AI berisiko menyebabkan kerugian yang tidak semestinya atau menunjukkan perilaku yang tidak diinginkan, sistem tersebut dapat diatasi, diperbaiki, dan/atau dinonaktifkan dengan aman sesuai kebutuhan.
  - Mekanisme juga harus tersedia, jika secara teknis memungkinkan, untuk memperkuat integritas informasi sambil memastikan penghormatan terhadap kebebasan berekspresi."
- **Akuntabilitas:** "Para pelaku AI harus bertanggung jawab atas berfungsinya sistem AI dengan baik dan atas penghormatan terhadap prinsip-prinsip di atas, berdasarkan peran mereka, konteks, dan konsisten dengan perkembangan terkini. Untuk tujuan ini, para pelaku AI harus memastikan ketertelusuran, termasuk yang berkaitan dengan kumpulan data, proses, dan keputusan yang dibuat selama siklus hidup sistem AI, untuk memungkinkan analisis keluaran dan respons sistem AI terhadap pertanyaan, yang sesuai dengan konteks dan konsisten dengan perkembangan terkini. Para pelaku AI, berdasarkan peran mereka, konteks, dan kemampuan mereka untuk bertindak, harus menerapkan pendekatan manajemen risiko sistematis pada setiap fase siklus hidup sistem AI secara berkelanjutan dan mengadopsi perilaku bisnis yang bertanggung jawab untuk mengatasi risiko yang terkait dengan sistem AI, termasuk, jika perlu, melalui kerja sama antara berbagai pelaku AI, pemasok pengetahuan AI dan sumber daya AI, pengguna sistem AI, dan pemangku kepentingan lainnya. Risiko tersebut meliputi risiko yang terkait dengan bias yang merugikan, hak asasi manusia termasuk keselamatan, keamanan, dan privasi, serta hak-hak buruh dan kekayaan intelektual."

Panduan Implementasi untuk Pembuat Kebijakan Pemerintah

Selain kelima prinsip ini, OECD menawarkan rekomendasi implementasi untuk pemerintah, yang meliputi:

- Mendorong ekosistem digital untuk AI melalui investasi dan infrastruktur
- Mempromosikan tenaga kerja terampil dan pendidikan inklusif
- Memastikan kepemimpinan sektor publik dalam adopsi AI yang dapat dipercaya
- Memfasilitasi kerja sama internasional dalam penelitian dan tata kelola AI

Relevansi bagi Para Profesional Tata Kelola AI

Para profesional tata kelola AI harus mempertimbangkan Prinsip-Prinsip OECD sebagai lensa dasar untuk pengembangan kebijakan. Prinsip-prinsip ini berfungsi sebagai ekspektasi dasar dalam kolaborasi lintas batas dan perjanjian multilateral, memberikan titik awal yang bermanfaat ketika membuat kebijakan etika AI internal atau membandingkan upaya kepatuhan dengan peraturan yang diantisipasi atau yang muncul.

7.2 KERANGKA KERJA MANAJEMEN RISIKO AI NIST (AI RMF)

Dengan munculnya praktik dan peraturan industri, memastikan penggunaan AI yang bertanggung jawab dan dapat dipercaya telah menjadi prioritas. Institut Standar dan Teknologi Nasional (NIST) merilis Kerangka Kerja Manajemen Risiko AI (AI RMF) untuk memandu organisasi dalam mengelola risiko yang terkait dengan sistem AI sepanjang siklus hidupnya. AI RMF, bersama dengan buku panduannya, adalah kerangka kerja terstruktur yang dibangun di sekitar empat fungsi inti: Tata Kelola, Pemetaan, Pengukuran, dan Pengelolaan. Kategori dan subkategori mendukung fungsi-fungsi ini untuk membantu organisasi mengatasi risiko khusus AI, termasuk keamanan, keadilan, privasi, penjelasan, dan ketahanan. Buku panduan AI RMF yang menyertainya menawarkan panduan implementasi praktis, dan organisasi dapat menyesuaikan "profil" dengan sektor, selera risiko, dan kasus penggunaan mereka. Kerangka kerja ini berfungsi sebagai alat pelengkap yang berharga untuk tata kelola AI berbasis risiko.

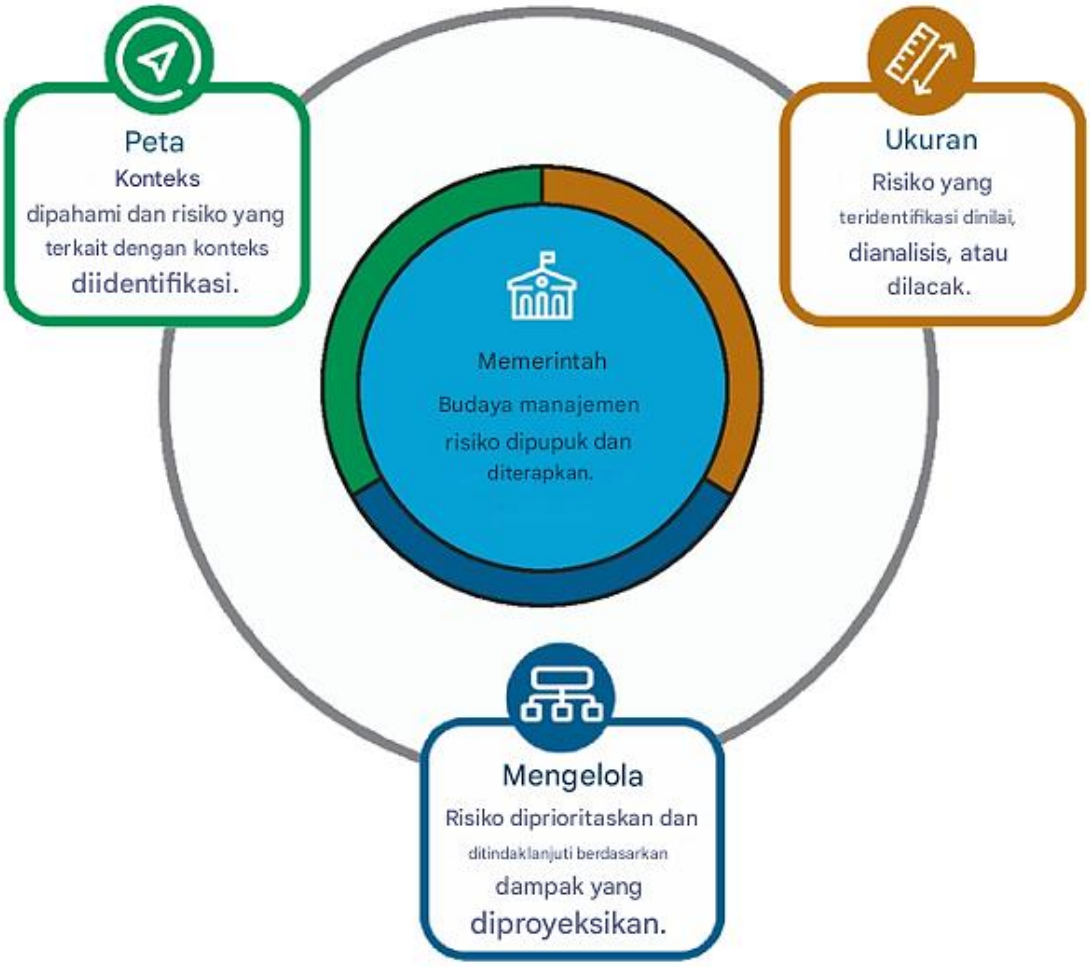
Organisasi yang membangun fungsi tata kelola AI dapat menggunakan AI RMF sebagai landasan mereka. Organisasi yang beroperasi di, atau menawarkan sistem AI ke, Uni Eropa tetap harus mematuhi Undang-Undang AI Uni Eropa, seperti yang dibahas dalam Bab 6.

Gambaran Umum NIST AI RMF

Diterbitkan pada Januari 2023, Kerangka Kerja Manajemen Risiko AI NIST (AI RMF) membahas kekhawatiran tentang konsekuensi yang tidak diinginkan dari penerapan AI. Tidak seperti kerangka kerja keamanan siber atau privasi tradisional, AI RMF berpusat pada karakteristik unik risiko AI, yang dapat bergantung pada konteks dan sulit untuk dikuantifikasi. Tujuan utama AI RMF adalah sebagai berikut:

- ✓ Memungkinkan AI yang dapat dipercaya dan bertanggung jawab melalui manajemen risiko yang terstruktur.
- ✓ Menyediakan kerangka kerja sukarela yang fleksibel dan sesuai untuk berbagai sektor dan tingkat kematangan.
- ✓ Mendorong kolaborasi di antara pengembang, penyebar, evaluator, dan pengguna AI.
- ✓ Mendukung inovasi sambil secara proaktif mengidentifikasi dan mengurangi dampak negatif.

AI RMF bertujuan untuk menetapkan bahasa dan metodologi yang terpadu untuk mengelola risiko terkait AI. Kerangka kerja ini mengambil inspirasi dari pengalaman luas NIST di bidang keamanan siber dan privasi, dengan mengadaptasi konsep-konsep dasar tersebut ke lanskap sosio-teknis AI yang kompleks. Fungsi-fungsi AI RMF digambarkan pada Gambar 7.1.



Gambar 7.1 Fungsi NIST AI RMF. Gambar dari Institut Standar dan Teknologi Nasional AS.

Apa yang Membuat Risiko AI Unik?
➢ Risiko muncul bukan hanya dari teknologi tetapi juga dari konteks. ➢ Banyak risiko AI berasal dari data yang bias, proksi, atau konteks penerapan yang memperkuat ketidaksetaraan. ➢ Perilaku yang muncul mungkin tidak dimaksudkan atau diperkirakan pada saat perancangan. ➢ Hasil mungkin tidak dapat diprediksi, karena algoritma yang buram atau adaptif. ➢ Kerugian dapat memengaruhi individu, kelompok, atau masyarakat, bahkan ketika kinerja tampak optimal. ➢ Kepatuhan terhadap undang-undang privasi menghadirkan tantangan unik. ➢ Tata kelola harus mempertimbangkan kualitas data, maksud desain, perilaku sistem, dan pengawasan manusia. ➢ Risiko AI meningkat dengan cepat. Setelah diterapkan, hasilnya dapat memengaruhi jutaan orang secara instan (misalnya, peringkat konten, keputusan otomatis).

Fungsi Tata Kelola: Membangun Fondasi untuk Manajemen Risiko AI

Fungsi Tata Kelola adalah landasan dari Kerangka Kerja Manajemen Risiko AI (AI RMF), dan sangat relevan bagi mereka yang bertanggung jawab untuk mengimplementasikan dan mengelola tata kelola AI. Fungsi ini menetapkan konteks organisasi, budaya, dan struktur yang diperlukan untuk mengelola risiko AI secara efektif. Tanpa tata kelola yang kuat, upaya untuk memetakan, mengukur, atau mengelola risiko mungkin terfragmentasi, tidak selaras, atau bahkan tidak terjadi sama sekali. Fungsi Tata Kelola mendukung akuntabilitas institusional, kesadaran risiko, dan kolaborasi lintas fungsi di seluruh siklus hidup AI.

Tata Kelola sangat penting dalam organisasi yang mengembangkan, menerapkan, atau menggunakan sistem AI. Fungsi ini meletakkan dasar untuk AI yang dapat dipercaya dengan menanamkan manajemen risiko ke dalam kebijakan dan praktik organisasi.

Rincian fungsi Tata Kelola dijelaskan di bagian berikut.

Kategori 1: Kebijakan, Proses, dan Prosedur

Organisasi harus mendefinisikan, menerapkan, dan memelihara kebijakan, prosedur, dan kontrol formal untuk memandu pengembangan, akuisisi, dan penggunaan sistem AI. Subkategori utama meliputi:

- Kebijakan dan prosedur manajemen risiko didokumentasikan.
- Prosedur didefinisikan yang mencakup pertimbangan khusus AI (misalnya, keadilan, kemampuan menjelaskan, ketahanan, otonomi privasi).
- Kebijakan yang mencerminkan nilai-nilai organisasi dan persyaratan hukum/regulasi ditetapkan.

Misalnya, organisasi perawatan kesehatan yang menerapkan alat AI diagnostik mengadopsi kebijakan yang mengharuskan setiap algoritma untuk menjalani audit keadilan pra-implementasi dan pemantauan penyimpangan pasca-implementasi.

Kategori 2: Peran dan Tanggung Jawab

Organisasi harus menetapkan peran dan akuntabilitas yang jelas di seluruh siklus hidup AI. Subkategori utama meliputi:

- Peran dan tanggung jawab untuk manajemen risiko didokumentasikan dan dikomunikasikan.
- Badan pengawas (misalnya, komite etika AI, audit internal) ditunjuk dan diberi wewenang untuk bertindak.
- Mekanisme eskalasi ditetapkan untuk risiko yang belum terselesaikan.

Model RACI dalam Tata Kelola AI

Banyak organisasi menggunakan model RACI (Bertanggung Jawab, Akuntabel, Dikonsultasikan, Diinformasikan) untuk menetapkan tanggung jawab. Misalnya, ilmuwan data mungkin bertanggung jawab atas pengembangan model, sementara tim hukum/kepatuhan bertanggung jawab untuk memastikan keselarasan dengan peraturan. Model RACI dibahas dalam Bab 2.

Kategori 3: Toleransi Risiko dan Selera Risiko

Organisasi harus mendefinisikan ambang batas toleransi risiko mereka untuk sistem AI dan menyelaraskannya dengan tujuan di seluruh perusahaan. Idealnya, organisasi dapat memperluas selera risiko dan toleransi risiko yang ada untuk memasukkan pertimbangan terkait AI. Subkategori utama meliputi:

- Pernyataan selera risiko dan toleransi risiko dikembangkan dan didokumentasikan.
- Keputusan tentang tingkat risiko yang dapat diterima didasarkan pada dampak bisnis dan konteks sosial, dan selaras dengan risiko dalam konteks lain.
- Masukan pemangku kepentingan dikumpulkan dari tim teknis, bisnis, dan kepatuhan saat menetapkan ambang batas.
- Tinjauan berkala terhadap toleransi risiko dilakukan seiring dengan perkembangan teknologi, peraturan, dan kasus penggunaan.

Misalnya, lembaga keuangan mendefinisikan toleransi risiko rendah untuk bias dalam persetujuan pinjaman tetapi toleransi yang lebih tinggi untuk eksperimen dalam alat produktivitas internal.

Kategori 4: Budaya Organisasi dan Komunikasi Internal

Budaya memainkan peran penting dalam keberhasilan tata kelola AI dan manajemen risiko. Tata kelola harus mempromosikan budaya kesadaran etis, komunikasi terbuka, dan kesadaran risiko. Subkategori utama meliputi:

- Para pemimpin menjadi teladan dan mempromosikan pengambilan keputusan yang berlandaskan risiko.
- Karyawan didorong untuk menyampaikan kekhawatiran terkait risiko AI tanpa takut akan pembalasan.
- Program pelatihan dan komunikasi memperkuat penggunaan AI yang etis.

Tata kelola AI harus selaras dengan bisnis dan, sejauh mungkin, terintegrasi ke dalam kerangka kerja dan struktur yang ada, sehingga meningkatkan kemungkinan keberhasilan dan menghasilkan hasil yang lebih baik dengan sistem AI.

Budaya Risiko AI vs. Budaya Keamanan

Meskipun keduanya menekankan kewaspadaan dan kesadaran risiko proaktif, budaya risiko AI juga membutuhkan pengkajian asumsi secara terus-menerus, perhatian pada konteks, dan penyertaan pertimbangan etis dan sosial—bukan hanya pengamanan teknis.

Fungsi Pemetaan: Memahami Konteks Sistem AI

Fungsi Pemetaan AI RMF membantu organisasi mengidentifikasi dan memahami karakteristik sistem AI dan konteks operasinya. Fungsi Pemetaan mencakup pemetaan komponen teknis, serta tujuan bisnis yang dimaksudkan dan yang dapat diprediksi, keterbatasan, dampak sosial, dan harapan pemangku kepentingan. Pemetaan merupakan aktivitas tata kelola yang penting karena risiko AI sangat kontekstual; sistem AI yang sama mungkin berisiko tinggi dalam satu kasus penggunaan dan berisiko rendah dalam kasus penggunaan lainnya. Oleh karena itu, pemetaan harus dilakukan secara iteratif di seluruh

siklus hidup AI untuk mencerminkan teknologi, peraturan, dan konteks penggunaan yang berkembang.

Fungsi Pemetaan membantu organisasi menghindari memperlakukan AI sebagai "kotak hitam." Sebaliknya, Pemetaan mendorong kejelasan, visibilitas, dan pemahaman, yang menginformasikan langkah-langkah selanjutnya dalam mengukur dan mengelola risiko.

Kategori 1: Konteks dan Tujuan yang Dimaksudkan

Ini melibatkan pemahaman mengapa dan bagaimana setiap instance sistem AI bersifat mendasar. Organisasi harus mendefinisikan dengan jelas tujuan sistem dan lingkungan tempat sistem tersebut beroperasi. Subkategori utama meliputi:

- Penggunaan, tujuan, dan keterbatasan yang dimaksudkan didokumentasikan.
- Penyalahgunaan yang dapat diprediksi dan aplikasi yang tidak disengaja dianalisis dan didokumentasikan.
- Konteks sosio-teknis (misalnya, faktor hukum, budaya, dan demografis) dianalisis.
- Keselarasan sistem AI dengan misi dan nilai-nilai dinilai.
- Dokumentasi dibuat dapat dilacak dan diaudit untuk tujuan tata kelola dan akuntabilitas.

Misalnya, sebuah lembaga transportasi umum yang memetakan perencana rute penumpang bertenaga AI mempertimbangkan geografi lokal, demografi komuter, dan tujuan kesetaraan layanan sebelum penerapan, untuk memastikan operasi yang aman, adil, dan etis.

Kategori 2: Desain dan Fungsionalitas Sistem

Seperti yang disebutkan di seluruh buku ini, Kategori 2 menegaskan bahwa organisasi harus mendokumentasikan cara kerja setiap sistem AI, termasuk input data, arsitektur model, dan logika pengambilan keputusan. Subkategori utama meliputi:

- ✓ Sumber data, langkah-langkah pra-pemrosesan, dan fitur model diidentifikasi.
- ✓ Desain interaksi manusia-AI dijelaskan.
- ✓ Metrik dan batasan kinerja model didefinisikan.

Menghindari Ketergantungan Berlebihan pada Metrik Kinerja

Kinerja dan kepercayaan adalah hal yang terpisah dan hubungannya sangat longgar (jika ada). Model dengan akurasi tinggi mungkin masih memperkuat stereotip yang berbahaya atau mengikis otonomi manusia. Pemetaan harus mencakup konteks kualitatif, bukan hanya kinerja kuantitatif.

Kategori 3: Pemangku Kepentingan dan Populasi yang Terdampak

Analisis pemangku kepentingan memastikan pemahaman tentang dampak langsung dan tidak langsung, khususnya bagi kelompok rentan atau yang kurang terwakili. Sistem AI harus memperlakukan orang yang terdampak secara adil. Subkategori utama meliputi:

- ➢ Pengguna utama dan pengambil keputusan diidentifikasi.
- ➢ Individu, komunitas, dan pihak ketiga yang terdampak dipertimbangkan.
- ➢ Proses keterlibatan pemangku kepentingan didokumentasikan.

Misalnya, ketika merancang algoritma penerimaan akademik, distrik sekolah mengumpulkan umpan balik dari beberapa pemangku kepentingan—termasuk siswa, orang tua, kelompok hak sipil, dan staf sekolah—untuk menilai potensi dampaknya.

Kategori 4: Identifikasi dan Prioritas Risiko

Pemetaan menghasilkan identifikasi potensi risiko, termasuk yang terkait dengan keselamatan, diskriminasi, privasi, manipulasi, dan dampak lingkungan.

Subkategori meliputi:

- Risiko yang diketahui dan dapat diprediksi didokumentasikan.
- Risiko diprioritaskan berdasarkan kemungkinan dan tingkat keparahannya.
- Sumber ketidakpastian dan kesenjangan pemahaman diidentifikasi.

Ketidakpastian Epistemik vs. Ketidakpastian Aleatorik
▪ Ketidakpastian epistemik mengacu pada hal-hal yang tidak diketahui dari pengetahuan yang terbatas (misalnya, data pelatihan yang tidak mencukupi). ▪ Ketidakpastian aleatorik mengacu pada keacakan yang melekat (misalnya, noise dalam data sensor).
Pemetaan harus mengidentifikasi kedua jenis risiko tersebut.

Fungsi Pengukuran: Menilai Risiko dan Kepercayaan AI

Seperti yang tersirat dalam istilahnya, fungsi Pengukuran RMF AI berfokus pada evaluasi seberapa baik sistem AI memenuhi harapan terkait kinerja, keandalan, kepercayaan, dan mitigasi risiko. Pengukuran menekankan evaluasi berkelanjutan menggunakan metode kuantitatif dan kualitatif. Pengukuran bukanlah tugas sekali saja, tetapi aktivitas berkelanjutan yang selaras dengan sifat dinamis dan berkembang dari sistem AI.

Fungsi Pengukuran membantu organisasi menentukan apakah setiap instance sistem AI beroperasi sesuai tujuan, dan apakah tetap berada dalam ambang risiko yang dapat diterima sepanjang siklus hidupnya.

Kategori 1: Metrik dan Indikator Risiko

Organisasi harus mendefinisikan metrik dan indikator untuk melacak aktivitas dan risiko terkait AI dari waktu ke waktu, sehingga memungkinkan mereka untuk memantau kemajuan dan mengidentifikasi potensi masalah. Metrik ini dapat mencakup dimensi teknis, operasional, etis, dan sosial. Subkategori utama:

- Kriteria pengukuran risiko ditetapkan untuk setiap risiko yang teridentifikasi.
- Metrik disesuaikan dengan konteks sistem dan pemangku kepentingan.
- Metrik diperbarui seiring dengan perkembangan pemahaman tentang perilaku sistem dan risiko.

Misalnya, algoritma pelacakan dan penyaringan pelamar kerja dipantau menggunakan metrik keadilan (misalnya, rasio dampak yang berbeda), akurasi berdasarkan kelompok demografis, dan survei kepuasan pemangku kepentingan.

Kategori 2: Metode Evaluasi

Evaluasi sistem AI harus mencakup berbagai metode, termasuk pengujian, simulasi, audit, dan tinjauan manusia, untuk mendapatkan pemahaman komprehensif tentang bagaimana sistem berperilaku. Subkategori meliputi:

- Metode pengujian (misalnya, pengujian unit, red teaming, blue teaming, dan stress testing) diterapkan.
- Metode kualitatif (misalnya, panel ahli, studi pengguna) melengkapi data kuantitatif.
- Audit pihak ketiga dipertimbangkan untuk kasus penggunaan berisiko tinggi.

Pentingnya Evaluasi Multimodal

Evaluasi keadilan, kemampuan menjelaskan, dan dampak negatif membutuhkan lebih dari sekadar angka. Wawasan kualitatif, seperti wawancara pemangku kepentingan, kelompok fokus, dan studi kasus, menambah kedalaman pemahaman tentang bagaimana sistem dapat menyebabkan dampak negatif atau menghasilkan hasil yang tidak terduga.

Kategori 3: Pemantauan Kinerja Sistem

Organisasi harus memantau perilaku sistem di lingkungan langsung untuk mendeteksi penurunan kinerja, penyimpangan, anomali, dan konsekuensi yang tidak diinginkan.
Subkategori meliputi:

- Pemantauan dan pencatatan kinerja berkelanjutan.
- Sistem pemantauan peristiwa mendeteksi kesalahan dan anomali.
- Siklus umpan balik memperbaiki dan menyesuaikan sistem.

Misalnya, sistem moderasi konten di platform media sosial menggunakan pemantauan waktu nyata untuk mendeteksi peningkatan kesalahan dan positif palsu setelah pembaruan sistem.

Kategori 4: Dokumentasi dan Pelaporan Risiko

Dokumentasi risiko yang transparan dan dapat ditindaklanjuti mendukung akuntabilitas dan membantu menghindari risiko yang tidak dapat diterima. Subkategori meliputi:

- ✓ Hasil pengukuran risiko didokumentasikan dan diarsipkan.
- ✓ Laporan tersedia untuk pemangku kepentingan dengan detail dan konteks yang sesuai.
- ✓ Hasilnya memberikan informasi untuk pembaruan sistem dan keputusan kebijakan di masa mendatang.

Kartu Model dan Kartu Sistem

Alat seperti kartu model dan kartu sistem (yang digunakan dalam model bahasa besar) dapat membantu mendokumentasikan kinerja, keterbatasan, dan risiko sistem AI dalam format yang terstruktur dan mudah diakses.

Fungsi Manajemen: Mengambil Tindakan untuk Mengatasi Risiko AI

Fungsi Manajemen AI RMF berfokus pada respons proaktif terhadap risiko dan peristiwa AI yang teridentifikasi melalui penyesuaian desain, kontrol operasional, intervensi kebijakan, dan keterlibatan pemangku kepentingan. Sementara fungsi Tata Kelola, Pemetaan, dan Pengukuran membangun pemahaman dan struktur, Manajemen mengoperasionalkan wawasan tersebut menjadi mitigasi risiko konkret dan peningkatan sistem.

Aktivitas manajemen harus berulang, peka terhadap konteks, dan responsif terhadap perubahan. Risiko dapat berkembang seiring dengan pergeseran data, perubahan perilaku pengguna, atau perkembangan harapan sosial, dan fungsi Manajemen membantu organisasi beradaptasi sesuai dengan hal tersebut.

Fungsi Manajemen dapat dianggap sebagai operasi sistem AI.

Kategori 1: Penanganan dan Mitigasi Risiko

Organisasi harus menerapkan dan memelihara kontrol (kebijakan, proses, atau prosedur) untuk mengurangi risiko ke tingkat yang dapat diterima berdasarkan konteks model AI dan toleransi risiko. Subkategori utama meliputi:

- Tindakan mitigasi diimplementasikan berdasarkan hasil penilaian risiko.
- Kontrol mencakup langkah-langkah teknis (misalnya, batasan algoritma) dan organisasi (misalnya, penegakan kebijakan).
- Penanganan risiko disesuaikan dengan kasus penggunaan sistem dan potensi bahayanya.

Menurut praktik manajemen risiko standar, semua keputusan penanganan risiko (baik penerimaan risiko, pengalihan risiko, penghindaran risiko, atau mitigasi risiko) harus disetujui oleh manajemen dan didokumentasikan.

Sebagai contoh, sistem pengenalan suara yang digunakan untuk perbankan berbasis telepon menerapkan verifikasi pembicara tambahan untuk pengguna dengan berbagai aksen untuk mengurangi bias tanpa mengurangi keamanan.

Kategori 2: Respons dan Pemulihan Insiden

Bahkan dengan kontrol terbaik sekalipun, upaya mitigasi tidak dapat menghilangkan semua risiko, artinya insiden dan pelanggaran masih dapat terjadi. Oleh karena itu, organisasi harus merencanakan insiden khusus AI dan menetapkan protokol yang jelas untuk mendeteksi, meningkatkan, dan menyelesaikannya. Subkategori utama meliputi:

- Buku panduan respons insiden khusus AI didokumentasikan dan dilatih.
- Respons insiden AI harus selaras dengan kerangka kerja keamanan siber dan respons insiden organisasi secara keseluruhan.
- Prosedur peningkatan dikaitkan dengan ambang risiko yang ditentukan.
- Pemulihan dan tinjauan pasca-tindakan mencakup siklus pembelajaran untuk mencegah terulangnya insiden.

Apa Itu "Insiden AI"?

Insiden AI dapat mencakup:

- Mobil otonom menyimpang dari jalur karena kegagalan model atau keadaan yang tidak terduga.
- Asisten AI memberikan nasihat yang berbahaya.
- Audit bias mengungkapkan perlakuan yang berbeda dalam keputusan-keputusan penting.

Organisasi harus memperlakukan insiden AI sebagai peluang pembelajaran, bukan hanya kegagalan kepatuhan atau operasional.

Kategori 3: Integrasi Manajemen Risiko

Manajemen risiko harus menjadi bagian integral dari proses organisasi rutin, bukan diperlakukan sebagai fungsi kepatuhan yang terisolasi saja. Subkategori utama meliputi:

- Risiko AI diintegrasikan dengan praktik keamanan siber, privasi, dan manajemen risiko perusahaan (ERM) yang lebih luas.
- Tim lintas fungsi mengoordinasikan keputusan risiko di seluruh departemen.
- Pengadaan, manajemen vendor, dan alat pihak ketiga mengikuti protokol risiko AI.

Misalnya, sebuah perusahaan menyelaraskan tata kelola siklus hidup model AI-nya dengan kerangka kerja manajemen risiko berbasis ISO 27001 yang ada, memungkinkan visibilitas risiko terpadu di seluruh domain keamanan siber, privasi, dan etika AI.

Kategori 4: Evolusi Risiko dan Penghentian Sistem

Lanskap risiko untuk sistem AI akan terus berkembang karena perubahan tujuan perusahaan, model bisnis, data, peraturan, dan nilai-nilai masyarakat. Organisasi harus mengevaluasi apakah sistem terus memberikan nilai, atau apakah sistem tersebut harus diperbarui, digunakan kembali, atau dihentikan. Subkategori utama meliputi:

- Sistem ditinjau secara berkala dan rutin untuk kesesuaian dan relevansi yang berkelanjutan.
- Perencanaan akhir masa pakai telah diterapkan untuk sistem AI dengan risiko atau keusangan yang diketahui.
- Penghentian penggunaan mencakup dokumentasi yang tepat dan penilaian dampak pasca-penggunaan.

Jangan Biarkan Model Bertahan Lebih Lama dari Tata Kelolanya

Sistem AI dapat bertahan lama setelah tujuan awalnya terpenuhi dan tim aslinya bubar. Tanpa perencanaan penghentian operasional, model yang sudah usang atau kurang diawasi dapat terus beroperasi secara sembunyi-sembunyi, menyebabkan kerugian yang tidak terpantau dan konsekuensi material.

Menyatukan Semuanya: Mengintegrasikan Kerangka Kerja Manajemen Risiko AI ke dalam Praktik Tata Kelola

Empat fungsi inti dari Kerangka Kerja Manajemen Risiko AI NIST (Mengatur, Memetakan, Mengukur, dan Mengelola) dimaksudkan untuk bersifat iteratif, saling

bergantung, dan adaptif. Namun, pemenuhan tujuan-tujuan ini bergantung pada organisasi. Tidak seperti alat kepatuhan linier atau berbasis daftar periksa, Kerangka Kerja Manajemen Risiko AI menekankan keselarasan berkelanjutan antara nilai-nilai organisasi, harapan pemangku kepentingan, dan perilaku sistem.

Ketika diterapkan secara efektif, Kerangka Kerja Manajemen Risiko AI membantu organisasi untuk:

- Membangun dan memungkinkan kesiapan organisasi untuk AI yang bertanggung jawab melalui struktur dan budaya tata kelola.
- Mencapai kesadaran situasional dengan memetakan sistem dan konteksnya, dan melalui pemantauan berkelanjutan.
- Melakukan penilaian berbasis bukti terhadap risiko, kinerja, dan kepercayaan.
- Menerapkan respons yang terarah dan disengaja terhadap peristiwa, insiden, ancaman, kegagalan, dan masalah etika.

Banyak organisasi menggunakan AI RMF untuk mengembangkan lapisan khusus AI di atas proses tata kelola dan manajemen risiko yang ada, selaras dengan model tata kelola seperti ISO/IEC 27001, manajemen risiko perusahaan (ERM), atau NIST CSF.

AI RMF dirancang untuk dapat diskalakan sesuai dengan kompleksitas sistem, ukuran organisasi, dan kematangan organisasi. Sebuah perusahaan rintisan kecil yang membangun chatbot dapat menerapkannya dalam bentuk yang ringan; perusahaan multinasional yang menerapkan AI di bidang jasa keuangan atau perawatan kesehatan dapat menerapkan semua kategori dan subkategori secara ketat, dengan beberapa lapisan pengawasan.

AI RMF vs. Kerangka Kerja Keamanan Siber NIST (CSF)

Meskipun dikembangkan secara terpisah, NIST AI RMF dan Kerangka Kerja Keamanan Siber NIST (CSF) memiliki asal usul yang sama. Keduanya merupakan kerangka kerja sukarela yang berfokus pada hasil yang dirancang untuk mempromosikan kepercayaan, ketahanan, dan manajemen risiko yang bertanggung jawab.

Tabel 7.1

Perbedaan Antara NIST AI RMF dan NIST CSF

Elemen	Saya punya RFM	CSF
Domain fokus	Siklus hidup sistem AI dan risiko sosio-teknis	Kontrol dan ketahanan keamanan siber
Jenis risiko yang ditangani	Bias, ketidaktransparan, penyalahgunaan, otonomi, penyimpangan, kerugian sosial	Kerahasiaan, integritas, ketersediaan (CIA), ketahanan operasional
Pengobatan ketidakpastian	Penanganan eksplisit terhadap ketidakpastian epistemik/aleatorik	Secara umum mengasumsikan adanya ancaman yang diketahui atau terukur.
Penekanan pada kemampuan menjelaskan dan keadilan	Tema inti	Tidak ditangani
Ketergantungan pada konteks sistem	Penting untuk pemetaan risiko	Kurang peka terhadap konteks, berbasis kontrol
Dimensi etis	Integral dengan kerangka kerja	Tidak dalam cakupan

Tabel 7.2	
Kapan Menggunakan AI RMF vs. CSF	
Kasus Penggunaan	**Kerangka yang Sesuai**
Mencegah malware dalam sistem TI	CSF
Mengevaluasi bias dalam algoritma penyaringan SDM.	AI RMF
Mengamankan infrastruktur cloud untuk beban kerja AI	CSF (dengan overlay khusus AI)
Mengatur etika dan dampak kendaraan otonom.	AI RMF (dengan konversi ke CSF untuk keamanan)

Namun, domain, asumsi, dan mekanisme mereka berbeda secara signifikan. Perbedaan tersebut digambarkan dalam Tabel 7.1. Tabel 7.2 mengilustrasikan berbagai kasus penggunaan di mana CSF atau AI RMF merupakan kerangka kerja yang tepat.

Pergeseran Filosofis Fundamental
CSF memperlakukan risiko sebagai sesuatu yang harus ditanggulangi dengan menggunakan kontrol yang sudah dikenal.
AI RMF memperlakukan risiko sebagai sesuatu yang muncul, sistemik, dan sarat nilai, yang membutuhkan strategi sosial, organisasi, dan teknis.

Dalam organisasi dengan program keamanan siber yang matang, AI RMF dapat ditambahkan ke CSF untuk memberikan pandangan holistik tentang risiko sistem AI, memadukan keamanan teknis dengan tata kelola risiko sosio-teknis.

7.3 NIST ARIA: EVALUASI MODEL AI BERLAPIS

Sistem AI semakin mampu dan digunakan untuk mendukung kasus penggunaan yang lebih penting dan kompleks, sehingga menciptakan kebutuhan akan mekanisme jaminan yang andal, berulang, dan transparan. Pada Mei 2024, NIST meluncurkan program *Assurance of AI Systems* (ARIA), sebuah inisiatif yang berorientasi ke masa depan yang dirancang untuk mengembangkan metode, alat, dan teknik pengukuran untuk mengevaluasi keamanan, kepercayaan, dan efektivitas sistem AI.

ARIA masih terus berkembang dan belum menjadi standar formal. Saat ini, ARIA berfokus pada risiko dan dampak sistem AI model bahasa besar (LLM). Pekerjaan di masa mendatang mungkin mencakup jenis AI generatif lainnya. Anda dapat menemukan informasi tentang ARIA di *https://ai-challenges.nist.gov/aria.*

Tujuan dan Posisi ARIA

Meskipun AI RMF dimaksudkan untuk bersifat sukarela dan fleksibel, ARIA mengambil pendekatan yang lebih terarah. Audiens utamanya adalah pemerintah federal AS, khususnya lembaga yang membangun atau menggunakan sistem AI untuk mendukung misi mereka. Namun, organisasi sektor swasta, lembaga penelitian, vendor, dan kontraktor juga dapat memperoleh manfaat dari ARIA sebagai cetak biru praktis untuk mengelola risiko AI di lingkungan yang diatur atau kompleks.

Alih-alih menggantikan AI RMF, ARIA melengkapinya melalui pendekatan berlapis untuk pengujian sistem AI. Dengan demikian, ARIA bertindak sebagai panduan untuk mengoperasionalkan bagian penilaian dari manajemen risiko AI.

Tiga Tingkat Pengujian

Evaluasi ARIA terhadap sistem AI mencakup tiga tingkat pengujian yang berbeda, masing-masing dengan fokus spesifik.

Pengujian Model

Tujuan pengujian model dalam ARIA adalah untuk mengkonfirmasi kemampuan yang diklaim oleh sistem AI. Menurut rencana evaluasi percontohan ARIA yang diterbitkan pada tahun 2024, pertanyaan pengujian model meliputi:

- ✓ Apakah sistem tersebut menunjukkan kemampuan yang dibutuhkan?
- ✓ Apakah sistem tersebut menunjukkan pengaman yang dibutuhkan?

Pengujian Tim Merah

Tujuan pengujian tim merah dalam ARIA adalah untuk menentukan apakah pengujian dapat menyebabkan sistem AI mengalami malfungsi atau berperilaku dengan cara yang tidak terduga atau tidak direncanakan. Mirip dengan pengujian tim merah dalam konteks lain (aplikasi, sistem informasi, jaringan), pengujian tim merah berupaya untuk mengelabui sistem AI agar melakukan tindakan yang tidak dirancang untuk dilakukannya. Sesuai dengan rencana evaluasi percontohan ARIA, pertanyaan pengujian meliputi:

- ➢ Bisakah sistem tersebut diinduksi untuk menghasilkan hasil yang melanggar?
- ➢ Dalam kondisi apa hasil yang melanggar terjadi?

Pengujian Lapangan

Tujuan pengujian lapangan dalam ARIA adalah untuk mengeksplorasi potensi dampak positif dan negatif dari sistem AI dalam penggunaan biasa oleh pengguna akhir. Rencana percontohan ARIA mencakup pertanyaan-pertanyaan seperti:

- ❖ Apakah pengguna terpapar informasi yang berdampak positif atau negatif selama penggunaan rutin?
- ❖ Apakah pengguna memahami informasi yang berdampak positif atau negatif yang disajikan kepada mereka?
- ❖ Berdasarkan paparan yang berdampak, apa tindakan selanjutnya yang dilakukan pengguna?

Penerapan Praktis ARIA

ARIA bukanlah daftar periksa, tetapi panduan yang memungkinkan organisasi untuk memprioritaskan tindakan berdasarkan misi, sumber daya, dan profil risiko mereka. Meskipun ARIA sendiri ada sebagai cetak biru untuk menguji sistem AI, hasil pengujian tersebut dapat memberikan informasi tentang aspek tata kelola AI yang lebih luas. Misalnya, jika pengujian model AI menghasilkan banyak pengecualian, para pemimpin bisnis dapat menyimpulkan perlunya fungsi tata kelola yang lebih menyeluruh atau efektif, seperti pengembangan persyaratan yang lebih baik.

ARIA dan Kepatuhan Cabang Eksekutif

Di pemerintahan federal AS, ARIA mendukung kepatuhan terhadap semakin banyak persyaratan federal seputar pengembangan dan penggunaan AI.

Secara khusus, ARIA membantu lembaga memenuhi harapan dari:

- **Memorandum OMB M-24-1O**, yang mewajibkan lembaga untuk melakukan penilaian risiko, memelihara inventaris AI, dan menyerahkan rencana tata kelola.
- **Perintah Eksekutif 1411O**, yang ditandatangani pada 30 Oktober 2023, yang mengarahkan lembaga federal untuk mempromosikan pengembangan teknologi AI yang aman, terjamin, dan dapat dipercaya. Perintah ini dicabut pada Januari 2025, yang mewakili pergeseran kebijakan yang signifikan.
- **Perintah eksekutif lainnya**, termasuk 14319 dan 14320.

Dengan menyelaraskan praktik evaluasi AI internal (seperti yang dipromosikan dalam ARIA) dengan mandat federal ini, lembaga dapat menyederhanakan pelaporan kepatuhan dan mendukung manajemen risiko yang kuat.

ARIA di Luar Pemerintah Federal

Meskipun awalnya dirancang dengan mempertimbangkan sektor publik AS, struktur dan pendekatan ARIA memiliki relevansi yang lebih luas. Perusahaan sektor swasta, LSM, dan lembaga akademis dapat mengadaptasi prinsip-prinsip ARIA untuk:

- Memperkuat program AI yang bertanggung jawab.
- Membangun kapasitas kelembagaan seputar risiko AI.
- Menyelaraskan kontrol internal dengan harapan peraturan eksternal.
- Melengkapi kerangka kerja tata kelola seperti NIST AI RMF atau ISO/IEC 42001.

Sifat modular ARIA membuatnya dapat diadaptasi secara luas, bahkan di luar konteks federal.

7.4 STANDAR ISO

Standar internasional memainkan peran penting dalam memungkinkan keselarasan global, interoperabilitas, dan kepercayaan pada teknologi yang ada dan yang sedang berkembang; AI tidak terkecuali. Di dalam Organisasi Internasional untuk Standardisasi (ISO) dan Komisi Elektroteknik Internasional (IEC), komite khusus yang terdiri dari para profesional berpengalaman telah aktif membangun dan membentuk lanskap standar AI untuk mendukung terminologi, metode, manajemen siklus hidup, dan praktik tata kelola yang konsisten. Dua standar yang paling relevan dan penting bagi para profesional tata kelola AI adalah ISO/IEC 22989 dan ISO/IEC 42001.

Bagian ini menjelaskan peran dan struktur dari kedua standar ISO AI ini. Bagian ini mengeksplorasi bagaimana 22989 memberikan definisi dan konsep dasar, sementara 42001 menjelaskan kerangka kerja sistem manajemen (mirip dengan ISO 27001 untuk keamanan informasi) bagi organisasi yang mengembangkan, menerapkan, atau menggunakan sistem AI secara bertanggung jawab. Daftar singkat standar ISO terkait menyusul setelah ini. Standar ISO dilisensikan kepada pengguna dan organisasi dan Anda dapat memperolehnya dari *https://www.iso.org/.*

ISO/IEC 22989: Konsep dan Terminologi untuk AI

Standardisasi dimulai dengan leksikon standar. Dalam percakapan profesional yang umum, sangat penting bahwa semua pihak memiliki pemahaman bersama tentang istilah-istilah seperti bias, penyimpangan, atau transparansi. ISO/IEC 22989 mendefinisikan konsep dan terminologi inti yang terkait dengan AI, menawarkan kosakata dasar untuk kebijakan, pengembangan, dan tata kelola. Standar ini tidak menentukan bagaimana AI harus dibangun atau dikelola. Sebaliknya, standar ini memastikan bahwa para pemangku kepentingan, termasuk pengembang, auditor, regulator, dan masyarakat, menggunakan istilah yang konsisten dan dapat dioperasikan saat membahas sistem AI.

Mengapa Terminologi Penting dalam Tata Kelola

Ketiadaan definisi bersama menciptakan miskomunikasi dan ketidakselarasan, terutama ketika sistem AI melintasi batas organisasi, yurisdiksi, atau budaya. ISO/IEC 22989 mendukung:

- Kolaborasi interdisipliner
- Kejelasan kontraktual untuk pengadaan AI dan perjanjian hukum
- Keselarasan peraturan lintas batas
- Kesiapan audit dan standar dokumentasi

Tidak Semua "AI" Sebenarnya Adalah AI

ISO 22989 menekankan ketelitian: sebuah "sistem AI" harus melibatkan unsur-unsur persepsi, penalaran, atau pembelajaran. Hal ini membedakan AI dari otomatisasi sederhana atau aturan bisnis, yang merupakan perbedaan penting untuk klasifikasi risiko dan ruang lingkup regulasi.

Definisi dan Konsep Utama dalam ISO 22989

Tabel 7.3 mencakup beberapa istilah inti terpilih yang didefinisikan dalam ISO 22989 yang sangat relevan untuk tata kelola AI.

Tabel 7.3

Definisi Tata Kelola AI Terpilih dalam ISO/IEC 22989

Ketentuan	**Definisi (Ringkasan)***
Kecerdasan Buatan (AI)	Kemampuan suatu sistem untuk melakukan tugas-tugas yang biasanya membutuhkan kecerdasan manusia.
Sistem AI	Suatu sistem rekayasa yang menghasilkan keluaran seperti prediksi, rekomendasi, atau keputusan, untuk serangkaian tujuan tertentu.
Mesin Pembelajaran (ML)	Sekumpulan metode yang memungkinkan sistem untuk meningkatkan kinerja pada suatu tugas berdasarkan data.
Otonomi	Kemampuan suatu sistem untuk beroperasi tanpa campur tangan manusia secara terus-menerus.
Transparansi	Sejauh mana cara kerja internal suatu sistem dan proses pengambilan keputusannya dapat dipahami.

Penjelasan	Sejauh mana manusia dapat memahami dan menafsirkan keluaran atau perilaku sistem AI.
Dapat dipercaya	Kemampuan sistem AI untuk beroperasi sesuai tujuan sambil meminimalkan risiko dan meningkatkan kepercayaan di antara para pemangku kepentingan.

Organisasi yang menggunakan istilah ISO 22989 dalam RFP dan kontrak lebih cenderung menghindari ambiguitas.

Kemampuan AI dan Perilaku Sistem

ISO/IEC 22989 menyediakan pengelompokan konseptual kemampuan AI, membantu membingkai diskusi tentang apa kemampuan sistem dan interaksinya dengan dunia. Ini termasuk:

- Persepsi (misalnya, pengenalan gambar)
- Penalaran (misalnya, optimasi, logika berbasis aturan)
- Pembelajaran (misalnya, jaringan saraf, pembelajaran terawasi)
- Interaksi (misalnya, bahasa alami, antarmuka gestur)
- Pengambilan keputusan (misalnya, perencanaan otonom, penilaian risiko)

Relevansi Siklus Hidup

Memahami kemampuan sistem AI membantu menentukan risiko mana yang relevan di setiap tahap siklus hidup, mulai dari pengumpulan data hingga penerapan antarmuka pengguna.

Tahapan Siklus Hidup Sistem AI

ISO/IEC 22989 mendefinisikan model siklus hidup untuk sistem AI dalam istilah konseptual tingkat tinggi. Namun, perlu diingat bahwa ISO/IEC 22989 adalah standar kosakata yang tidak memberikan informasi tentang implementasi siklus hidup. Kerangka kerja siklus hidup yang lebih detail, seperti ISO/IEC 23053 (siklus hidup pembelajaran mesin) dan ISO/IEC 5338 (manajemen siklus hidup AI), menjelaskan tahapan seperti berikut:

- ✓ Persyaratan bisnis dan desain
- ✓ Akuisisi dan pemrosesan data
- ✓ Pengembangan dan pelatihan model
- ✓ Integrasi dan pengujian sistem
- ✓ Penyebaran dan pengoperasian
- ✓ Pemantauan, pemeliharaan, dan penghentian

Contoh perusahaan ritel menyelaraskan proses pengembangan AI-nya dengan tahapan siklus hidup ISO/IEC 22989, memungkinkan auditor untuk mengetahui di mana pengawasan manusia diperlukan selama desain, pelatihan, penyebaran, dan pengoperasian.

ISO/IEC 42001: Sistem Manajemen AI

ISO/IEC 42001, yang diterbitkan pada tahun 2023, adalah sistem manajemen bersertifikasi pertama di dunia yang dirancang khusus untuk organisasi yang mengembangkan, menyediakan, atau menggunakan sistem AI. Dimodelkan berdasarkan standar yang sudah

mapan seperti ISO/IEC 27001 (keamanan informasi) dan ISO 9001 (kualitas), ISO/IEC 42001 membantu organisasi mengoperasionalkan tata kelola AI dengan menanamkan kepercayaan, transparansi, dan pengendalian risiko ke dalam praktik sehari-hari mereka.

Di mana ISO/IEC 22989 menyediakan terminologi umum, ISO/IEC 42001 mendefinisikan sistem tata kelola terstruktur, lengkap dengan peran, kebijakan, dokumentasi, audit internal, dan siklus peningkatan berkelanjutan, yang memungkinkan organisasi untuk menunjukkan penggunaan AI yang bertanggung jawab melalui sertifikasi formal. Seperti standar ISO lainnya, ISO/IEC 42001 dapat diskalakan dan fleksibel, sama cocoknya untuk perusahaan rintisan dan organisasi manufaktur multinasional. ISO/IEC 42001 dan 22989 saling berkaitan erat.

Tujuan dan Ruang Lingkup ISO/IEC 42001

ISO/IEC 42001 dapat digunakan oleh organisasi mana pun yang terlibat dalam rantai nilai AI, termasuk pengembang AI (perancang model, ilmuwan data), penyebar AI (penyedia platform, integrator), pengguna AI (perusahaan yang menggunakan alat AI pihak ketiga), dan bahkan organisasi yang secara tidak langsung terpengaruh oleh AI (mereka yang menggunakan fitur AI tertanam).

ISO/IEC 42001 tidak menentukan cara membangun model AI. Standar ini mendefinisikan bagaimana organisasi mengelola sistem AI, melalui kepemimpinan, kebijakan, kontrol, peran, dan proses tata kelola untuk seluruh siklus hidup sistem AI.

Persyaratan Inti ISO/IEC 42001

ISO/IEC 42001 mengikuti struktur tingkat tinggi yang sama dengan standar manajemen ISO lainnya, sehingga kompatibel dengan ISO 27001, 31000, dan lainnya. Komponen-komponen kunci meliputi:

- Konteks organisasi: Memahami faktor internal dan eksternal yang memengaruhi penggunaan AI, termasuk hukum dan standar yang berlaku
- Kepemimpinan dan komitmen: Manajemen puncak menetapkan arah, mengalokasikan sumber daya, dan memikul tanggung jawab
- Perencanaan: Penilaian risiko, tujuan, dan tindakan untuk mencapai tujuan kepercayaan
- Dukungan: Sumber daya, kompetensi, kesadaran, dan informasi terdokumentasi
- Operasi: Proses desain, akuisisi, pengembangan, dan penerapan
- Evaluasi kinerja: Pemantauan, audit, dan tinjauan manajemen
- Peningkatan: Tindakan korektif dan mekanisme peningkatan berkelanjutan

Manajemen Risiko dalam ISO/IEC 42001

Pemikiran berbasis risiko merupakan inti dari 42001. Menurut standar ini, organisasi harus mengidentifikasi dan menilai risiko yang terkait dengan sistem AI, mendefinisikan kriteria untuk tingkat risiko yang dapat diterima (misalnya, selera risiko dan toleransi risiko), dan menerapkan kontrol, mitigasi, dan prosedur pemantauan untuk memastikan bahwa risiko tetap berada dalam batas yang dapat diterima.

Jenis sumber risiko yang disebutkan oleh ISO/IEC 42001 meliputi:

- Bias algoritmik atau hasil yang tidak adil

- ❖ Perilaku yang tidak diinginkan dari model pembelajaran mesin
- ❖ Ketergantungan berlebihan pada keluaran AI tanpa pengawasan manusia
- ❖ Ketidakpatuhan terhadap peraturan atau kerusakan reputasi

Keselarasan dengan AI RMF

Persyaratan risiko ISO/IEC 42001 selaras dengan fungsi NIST AI RMF (Mengatur, Memetakan, Mengukur, Mengelola), menyediakan jembatan antara strategi dan sertifikasi formal. Kedua standar tersebut merupakan sumber yang sangat baik untuk struktur tata kelola AI dan proses terkait.

Dimensi Kepercayaan dalam 42001

Sesuai standar ISO/IEC 42001, organisasi diharuskan mempertimbangkan beberapa faktor kepercayaan dalam praktik manajemen AI mereka, termasuk:

- Transparansi dan penjelasan
- Keadilan dan non-diskriminasi
- Ketangguhan dan daya tahan
- Akuntabilitas dan pengawasan manusia
- Privasi dan tata kelola data

Anda dapat menganggap ini sebagai ukuran toleransi risiko. Pertimbangan ini memengaruhi rencana penanganan risiko, prinsip desain sistem, dan strategi keterlibatan pemangku kepentingan. Misalnya, sebuah perusahaan jasa keuangan memasukkan keadilan dan transparansi sebagai tujuan kepercayaan yang didokumentasikan dalam sistem manajemen AI 42001-nya, yang memicu tindakan mitigasi setiap kali ukuran utama berada di bawah ambang batas yang ditentukan.

Menerapkan ISO/IEC 42001 dalam Praktik

Menerapkan ISO/IEC 42001 bukan hanya tentang mendapatkan sertifikasi atau mencentang kotak; ini tentang mencapai tingkat kualitas dan kinerja yang lebih tinggi. Selain itu, ini tentang menanamkan disiplin tata kelola ke dalam proses organisasi untuk merancang, membangun, dan mengelola sistem AI. Bagian ini menguraikan pendekatan untuk sertifikasi, termasuk analisis kesenjangan, keselarasan dengan standar ISO lainnya, kesiapan untuk audit internal, dan penskalaan tata kelola di berbagai tim dan sistem.

Melakukan Penilaian Kesiapan

Sebelum implementasi penuh ISO/IEC 42001, organisasi harus melakukan analisis kesenjangan untuk membandingkan praktik saat ini dengan persyaratan ISO/IEC 42001. Penilaian ini meliputi:

- Inventaris sistem AI dan kasus penggunaan
- Inventaris kebijakan dan prosedur saat ini yang mengatur pengembangan dan penggunaan AI
- Inventaris peran dan akuntabilitas yang ada dengan pengawasan AI
- Inventaris dokumentasi internal, pelatihan, dan protokol pengujian

Misalnya, sebuah jaringan ritel yang melakukan penilaian kesenjangan ISO 42001 menemukan bahwa mereka memiliki proses pengembangan model yang kuat tetapi tidak ada pengawasan formal terhadap solusi AI yang disediakan vendor. Kesenjangan ini menyoroti kebutuhan untuk memperluas proses manajemen risiko pihak ketiga ke vendor AI pihak ketiga.

Integrasi dengan Sistem Manajemen yang Ada

Karena ISO/IEC 42001 mengikuti struktur tingkat tinggi yang sama dengan standar tata kelola ISO lainnya, standar ini dapat diintegrasikan dengan proses yang terinspirasi ISO yang sudah ada, termasuk:

- ISO/IEC 27001 (Manajemen Keamanan Informasi)
- ISO 9001 (Manajemen Mutu)
- ISO/IEC 27701 (Manajemen Privasi Informasi)

Kontrol Bersama di Seluruh Sistem

Satu register risiko, rencana respons insiden, atau fungsi audit seringkali dapat mendukung berbagai standar. ISO/IEC 42001 mendorong penggunaan kembali sumber daya jika sesuai. Mengapa harus memiliki banyak proses terpisah dan terisolasi ketika menggabungkannya menjadi satu proses akan lebih efektif dan efisien?

Mengembangkan Kebijakan dan Prosedur Khusus AI

Saat organisasi mengembangkan proses tata kelola AI, dokumen-dokumen kunci yang mendukung implementasi meliputi kebijakan manajemen siklus hidup sistem AI, prinsip-prinsip AI etis yang selaras dengan nilai-nilai organisasi, prosedur untuk deteksi dan mitigasi bias, persyaratan pencatatan dan pemantauan untuk keluaran sistem AI, dan protokol keterlibatan pemangku kepentingan.

Dokumen-dokumen ini harus mencerminkan konteks organisasi, termasuk risiko spesifik sektor, kewajiban peraturan, dan harapan pemangku kepentingan. Contoh penyedia layanan kesehatan yang menerapkan 42001 adalah dengan membentuk komite pengawasan AI lintas fungsi dan mengembangkan kebijakan baru yang mensyaratkan validasi algoritma sebelum digunakan dalam pengaturan klinis.

Pelatihan, Kesadaran, dan Budaya Organisasi

ISO/IEC 42001 sangat menekankan kompetensi dan kesadaran. Implementasi yang sukses membutuhkan staf teknis untuk memahami kewajiban tata kelola, pengguna bisnis dan manajer mengetahui cara mengenali dan meningkatkan risiko terkait AI, dan kepemimpinan mencontohkan budaya penggunaan AI yang etis dan bertanggung jawab.

Melampaui Kepatuhan

Sertifikasi bukanlah tujuan akhir, melainkan indikator kematangan. Organisasi yang menerapkan ISO/IEC 42001 sebagai disiplin tata kelola yang dinamis akan membangun sistem AI yang lebih mudah beradaptasi, tangguh, dan tepercaya dari waktu ke waktu.

Standar ISO Lain yang Relevan untuk Tata Kelola AI

Meskipun ISO/IEC 22989 dan ISO/IEC 42001 berfungsi sebagai standar dasar untuk tata kelola AI, beberapa standar ISO dan IEC tambahan mendukung disiplin ilmu spesifik tentang desain, penerapan, dan pengawasan AI yang bertanggung jawab. Standar tambahan ini membantu organisasi meningkatkan program tata kelola mereka dengan menangani domain spesifik kepercayaan, praktik operasional, dan kategori risiko.

ISO/IEC 23894: Manajemen Risiko AI

Standar ini melengkapi ISO/IEC 42001 dengan menyediakan kerangka kerja khusus untuk manajemen risiko dalam sistem AI. Standar ini selaras dengan ISO 31000 (Manajemen Risiko Perusahaan) dan menawarkan panduan yang disesuaikan tentang identifikasi dan analisis risiko spesifik AI, penentuan kontrol yang tepat, dan peninjauan postur risiko sepanjang siklus hidup AI.

Perlu dicatat bahwa ISO/IEC 23894 bukan tentang manajemen risiko menggunakan AI, tetapi lebih tentang manajemen risiko AI. Standar ini membantu organisasi mengoperasionalkan penilaian risiko dalam desain dan penerapan sistem AI, terutama ketika diintegrasikan dengan 42001.

ISO/IEC 38507: Tata Kelola TI (Implikasi Tata Kelola AI)

Sebagai bagian dari seri ISO 38500 tentang tata kelola TI, standar ini memberikan panduan tingkat strategis bagi anggota dewan dan eksekutif tentang pengawasan penggunaan AI. Standar ini membahas peran dan tanggung jawab untuk pengawasan AI, selera risiko, dan keselarasan etika, ditambah akuntabilitas dan transparansi di tingkat kepemimpinan. Standar ini menjembatani kesenjangan antara pengawasan eksekutif dan kontrol operasional, sehingga bermanfaat untuk pengarahan dewan dan perencanaan komite tata kelola.

ISO/IEC TR 24028: Kepercayaan dalam AI

Standar ini adalah laporan teknis (bukan standar yang dapat disertifikasi) yang memperkenalkan konsep kepercayaan AI dan menjelaskan mekanisme untuk mencapainya. Topik meliputi transparansi, keandalan, ketahanan, privasi, dan keadilan.

ISO/IEC TR 24028 memberikan landasan konseptual untuk klaim kepercayaan dan, bersama dengan manajemen risiko, dapat memberikan informasi untuk desain kontrol dalam implementasi ISO/IEC 42001.

ISO/IEC 42005: Kecerdasan Buatan (AI) (Penilaian Dampak Sistem AI)

Standar ini mendefinisikan metodologi untuk menilai dampak sistem AI —dan aplikasi yang dapat diprediksi—pada individu, kelompok, dan masyarakat secara keseluruhan. Penilaian dampak semakin dibutuhkan berdasarkan undang-undang seperti Undang-Undang AI Uni Eropa. ISO/IEC 42005 dapat menjadi metode standar untuk melakukan dan mendokumentasikan evaluasi ini.

Standar AI ISO dalam Lanskap Tata Kelola Global

Standar ISO merupakan komponen vital dari tata kelola AI global. Seiring munculnya regulasi nasional dan regional—dari Undang-Undang AI Uni Eropa hingga Perintah Eksekutif dan regulasi AS tentang AI—ISO menawarkan landasan non-regulasi yang terharmonisasi secara global yang mendukung:

- Kosakata dan metodologi standar
- Konsistensi lintas batas
- Adaptasi spesifik sektor
- Implementasi netral vendor
- Kesiapan sertifikasi dan audit

Organisasi yang selaras dengan ISO/IEC 22989 dan 42001 memposisikan diri untuk memenuhi harapan hukum, etika, operasional, dan sosial yang muncul sambil menunjukkan kepemimpinan dalam AI yang dapat dipercaya.

ISO vs. Kepatuhan Hukum

Kepatuhan ISO bersifat sukarela tetapi sering berfungsi sebagai bukti uji tuntas, kematangan risiko, dan keselarasan peraturan. Namun, kepatuhan dan sertifikasi ISO sering dikutip dalam perjanjian hukum, memungkinkan organisasi untuk meminta pertanggungjawaban pihak lain terhadap standar internasional ini.

Standar ISO tidak menghambat inovasi; sebaliknya, standar tersebut menciptakan batasan pragmatis yang memungkinkan inovasi yang aman, penerapan yang adil, dan akuntabilitas yang transparan. Seiring dengan semakin matangnya badan pengatur dan ekosistem sertifikasi, standar ISO kemungkinan akan menjadi:

- ✓ Tolok ukur untuk penilaian kesesuaian AI
- ✓ Model referensi dalam panduan penegakan hukum
- ✓ Blok bangunan untuk kode praktik khusus sektor

7.5 STANDAR TERKAIT AI LAINNYA

Meskipun inisiatif OECD, NIST, dan ISO membentuk inti dari standar tata kelola AI saat ini, kerangka kerja berpengaruh lainnya sedang muncul. Ini termasuk peraturan khusus wilayah, pedoman yang dipimpin industri, dan kerangka kerja multi-pemangku kepentingan yang mendukung pengembangan dan penerapan AI yang bertanggung jawab. Bagian ini mengeksplorasi beberapa di antaranya, menekankan pengaruh, kontribusi unik, dan relevansinya bagi para profesional tata kelola AI.

Pedoman Etika Kelompok Pakar Tingkat Tinggi Uni Eropa (HLEG)

Sebelum Undang-Undang AI Uni Eropa, Komisi Eropa membentuk kelompok pakar untuk mendefinisikan prinsip-prinsip etika untuk AI yang dapat dipercaya. Pada tahun 2019, Kelompok Pakar Tingkat Tinggi (HLEG) tentang AI merilis pedoman etika untuk AI yang dapat dipercaya. Komponen utamanya meliputi:

- ➢ Tujuh persyaratan inti untuk AI yang dapat dipercaya:
    1. Peran dan pengawasan manusia
    2. Ketangguhan dan keamanan teknis
    3. Privasi dan tata kelola data
    4. Transparansi

5. Keragaman, non-diskriminasi, dan keadilan
6. Kesejahteraan masyarakat dan lingkungan
7. Akuntabilitas

- Empat prinsip etika untuk AI yang dapat dipercaya:
  1. Penghormatan terhadap otonomi manusia
  2. Pencegahan bahaya
  3. Keadilan
  4. Kejelasan
- Daftar periksa penilaian mandiri bagi organisasi untuk mengevaluasi kepatuhan

Pedoman HLEG secara langsung memengaruhi Undang-Undang AI Uni Eropa, khususnya dalam kategorisasi risiko dan kewajiban untuk sistem berisiko tinggi.

Proses Hiroshima G7 dan Kode Etik

Proses Hiroshima G7 diluncurkan pada tahun 2023 sebagai respons terhadap evolusi pesat AI generatif. Pada Oktober 2023, G7 menerbitkan Prinsip Panduan dan Kode Etik untuk Sistem AI Tingkat Lanjut.

Prinsip Panduan dan Kode Etik untuk organisasi yang mengembangkan sistem AI canggih berfokus pada poin-poin berikut:

1. Mengidentifikasi, mengevaluasi, dan mengurangi risiko di seluruh siklus hidup AI.
2. Memperbaiki pola penyalahgunaan, setelah penerapan, termasuk penempatan di pasar.
3. Melaporkan secara publik kemampuan, keterbatasan, dan domain penggunaan yang tepat dan tidak tepat dari sistem AI canggih.
4. Berupaya menuju berbagi informasi dan pelaporan insiden yang bertanggung jawab di antara organisasi yang mengembangkan sistem AI canggih, termasuk dengan industri, pemerintah, masyarakat sipil, dan akademisi.
5. Mengembangkan, menerapkan, dan mengungkapkan kebijakan tata kelola dan manajemen risiko AI.
6. Berinvestasi dan menerapkan kontrol keamanan yang kuat.
7. Mengembangkan dan menerapkan mekanisme otentikasi dan asal usul konten yang andal.
8. Memprioritaskan penelitian untuk mengurangi risiko sosial, keselamatan, dan keamanan serta memprioritaskan investasi dalam langkah-langkah mitigasi yang efektif.
9. Memprioritaskan pengembangan sistem AI canggih untuk mengatasi tantangan terbesar di dunia.
10. Memajukan pengembangan dan, jika sesuai, adopsi standar teknis internasional.
11. Menerapkan langkah-langkah input data dan perlindungan yang tepat untuk data pribadi. dan kekayaan intelektual."

Artefak-artefak ini dan lainnya tersedia di *https://g7g20-documents.org/;* cari Kode Etik AI dan Prinsip Panduan AI.

Kerangka Kerja Desain yang Selaras secara Etis IEEE

Institut Insinyur Listrik dan Elektronik (IEEE) memiliki peran penting dalam membentuk fondasi etika desain sistem AI melalui inisiatif Desain yang Selaras secara Etis (EAD) dan seri standar IEEE 7000. Standar IEEE yang relevan adalah:

- **IEEE 7OOO:** Proses model untuk mengatasi masalah etika dalam desain sistem
- **IEEE 7OO1:** Transparansi dalam sistem otonom
- **IEEE 7OO2:** Privasi data dalam sistem AI
- **IEEE 7OOC:** Pertimbangan bias algoritmik
- **IEEE 7O1O:** Metrik kesejahteraan untuk sistem otonom

Seri IEEE 7000 menganggap etika sebagai pertimbangan desain integral, yang dimasukkan ke dalam analisis persyaratan, pemodelan sistem, dan validasi.

Anda dapat menemukan sumber daya ini di *https://standards.ieee.org/.*

Panduan Tata Kelola AI Forum Ekonomi Dunia (WEF)

Forum Ekonomi Dunia (WEF) telah menerbitkan panduan dan buku pedoman yang berfokus pada tata kelola AI untuk para pembuat kebijakan dan pemimpin organisasi. Kontribusi utama meliputi:

- Panduan Aliansi Tata Kelola AI (202C): Panduan tata kelola lintas sektor untuk penerapan AI yang bertanggung jawab
- Rekomendasi Presidio tentang AI yang Bertanggung Jawab (202O): Model multi-pemangku kepentingan untuk tata kelola berbasis risiko
- Panduan ChatGPT dan AI Generatif untuk Pembuat Kebijakan (202C): Pemetaan risiko dan panduan desain regulasi

Anda dapat memperoleh sumber daya ini dari *https://initiatives.weforum.org/ai-governance-alliance/home*.

Kerangka Kerja OECD untuk Klasifikasi Sistem AI

Terpisah dari prinsip-prinsip etika yang dibahas sebelumnya dalam bab ini, OECD merilis kerangka kerja untuk klasifikasi sistem AI pada tahun 2022. Kerangka kerja ini mengkategorikan sistem AI berdasarkan konteks, otonomi, interaksi manusia, skala dampak, dan pihak-pihak yang terpengaruh. Meskipun mirip dengan klasifikasi risiko dalam Undang-Undang AI Uni Eropa, kerangka kerja klasifikasi OECD bukanlah standar yang mengikat.

Rekomendasi UNESCO Perserikatan Bangsa-Bangsa tentang Etika AI

Pada tahun 2021, UNESCO mengadopsi standar global pertama tentang etika AI, yang berlaku untuk semua 194 negara anggota. Saat ini, standar ini bukanlah peraturan yang mengikat, melainkan praktik yang diusulkan. Rekomendasi UNESCO tentang etika AI mempromosikan desain yang berpusat pada hak asasi manusia dan mencakup:

- Penilaian dampak etis
- Tata kelola dan pengawasan etis
- Kebijakan data
- Keberlanjutan lingkungan
- Kesetaraan gender
- Perlindungan keragaman budaya

- Transparansi algoritma
- Kesehatan dan kesejahteraan sosial

Anda dapat menemukan isi lengkap dokumen UNESCO di *https://unesdoc.unesco.org/ark:/48223/pf0000381137.*

Kerangka Kerja Tata Kelola AI Model Singapura

Singapura telah muncul sebagai pemimpin dalam tata kelola AI berbasis model dan netral sektor. Kerangka Kerja Tata Kelola AI Modelnya, yang dirilis oleh Otoritas Pengembangan Media Infokom Singapura (IMDA) pada Mei 2024, menekankan:

- Akuntabilitas
- Kualitas data
- Pengembangan dan penerapan yang terpercaya
- Pelaporan insiden
- Pengujian dan jaminan
- Keamanan
- Asal usul konten
- Keamanan dan keselarasan R&D
- AI untuk kepentingan publik

Anda dapat menemukan kerangka kerja ini di *https://aiverifyfoundation.sg/resources/mgf-gen-ai/.*

Peta Jalan ISO/IEC JTC 1/SC 42 dan Standar Masa Depan

Subkomite ISO yang bertanggung jawab atas AI, yang dikenal sebagai ISO/IEC JTC 1/SC 42, telah merilis peta jalan untuk standar AI di masa mendatang. Area standardisasi yang akan datang meliputi:

- Penilaian dampak sistem AI
- Evaluasi kualitas dan kinerja AI
- Metodologi mitigasi bias
- Jaminan siklus hidup AI

Semua area ini memiliki praktik yang sudah ada saat ini, tetapi perlu ditingkatkan melalui pengalaman industri dan inovasi AI baru.

Para profesional tata kelola AI harus memantau peta jalan SC 42, karena standar yang akan datang kemungkinan akan memengaruhi skema sertifikasi di masa mendatang, kontrak pengadaan, dan kepatuhan lintas batas.

7.6 RINGKASAN

Sistem AI bukan lagi alat eksperimental yang beroperasi secara terisolasi; sistem ini tertanam di berbagai industri dan fungsi masyarakat, membawa janji besar sekaligus risiko yang signifikan. Dalam konteks ini, semakin banyak standar dan kerangka kerja AI yang muncul untuk memandu organisasi dalam membangun, menerapkan, dan mengelola AI secara bertanggung jawab. Bab ini memperkenalkan upaya internasional dan nasional utama untuk menyediakan pendekatan terstruktur, interoperabel, dan tepercaya terhadap tata kelola AI.

Prinsip-prinsip AI OECD meletakkan dasar etika bagi sebagian besar aktivitas kebijakan AI di dunia. Dengan mengartikulasikan lima prinsip inti (termasuk nilai-nilai yang berpusat pada manusia, transparansi, dan akuntabilitas), OECD menciptakan kerangka kerja yang fleksibel dan tidak mengikat yang telah memengaruhi sistem regulasi formal, seperti Undang-Undang AI Uni Eropa. Prinsip-prinsip ini berfungsi sebagai kompas moral dan strategis, memberikan visi bersama tentang apa yang harus dicapai oleh AI yang bertanggung jawab, bahkan ketika teknologi dan konteks berkembang.

Di Amerika Serikat, *National Institute of Standards and Technology* (NIST) telah mengembangkan Kerangka Kerja Manajemen Risiko AI (*AI Risk Management Framework*/AI RMF), alat sukarela tetapi komprehensif untuk membantu organisasi mengelola risiko khusus AI. Distrukturkan di sekitar empat fungsi inti (Mengatur, Memetakan, Mengukur, dan Mengelola), AI RMF mendorong organisasi untuk mengambil pandangan siklus hidup yang matang terhadap sistem AI (mirip dengan SDLC), dengan mengakui vektor risiko teknis dan sosio-teknis. Kerangka kerja ini cukup fleksibel untuk diterapkan di berbagai industri, namun cukup ketat untuk mendukung fungsi penjaminan dan audit.

Berdasarkan AI RMF, program ARIA (*Assessing Risks and Impacts of AI*) dari NIST menangani pertanyaan yang berbeda: bagaimana kita dapat memverifikasi bahwa sistem AI benar-benar memenuhi tujuan keselamatan, keadilan, dan kepercayaan yang dimaksudkan? ARIA menyediakan metodologi, alat, dan metrik yang berorientasi pada tindakan yang mendukung pengujian, pemantauan, dan validasi perilaku AI dalam kondisi dunia nyata. ARIA menekankan tidak hanya metrik kinerja tetapi juga kemampuan menjelaskan, ketahanan, dan dinamika interaksi manusia-AI, memungkinkan operasionalisasi prinsip-prinsip AI yang dapat dipercaya secara lebih mendalam.

Secara internasional, ISO telah mengembangkan beberapa standar AI yang memberikan struktur, terminologi, dan jalur sertifikasi. ISO/IEC 22989 menawarkan definisi dasar tentang konsep, sistem, dan tahapan siklus hidup AI, yang mendorong konsistensi di seluruh domain teknis dan regulasi. Sebaliknya, ISO/IEC 42001 memperkenalkan sistem manajemen yang dapat disertifikasi untuk AI, membantu organisasi menanamkan praktik tata kelola, pengendalian risiko, dan mekanisme peningkatan berkelanjutan ke dalam operasi sehari-hari mereka. Standar ISO tambahan lebih lanjut mendukung manajemen risiko, penilaian dampak, dan kepercayaan.

Beberapa kerangka kerja global lainnya memainkan peran pendukung yang vital. Pedoman Etika Kelompok Pakar Tingkat Tinggi Uni Eropa, Proses Hiroshima G7, seri IEEE 7000, rekomendasi etika UNESCO, Kerangka Kerja Tata Kelola AI Model Singapura, dan alat-alat dari Forum Ekonomi Dunia masing-masing berkontribusi pada mosaik AI yang aman, etis, dan bertanggung jawab yang lebih luas. Meskipun beragam dalam cakupannya, mereka memiliki tema umum akuntabilitas, keadilan, etika, transparansi, dan desain yang berpusat pada manusia. Bersama-sama, standar dan kerangka kerja ini memungkinkan para profesional tata kelola AI untuk membangun sistem yang tidak hanya patuh tetapi juga selaras secara etis, responsif secara sosial, dan tangguh dalam menghadapi tantangan di masa depan.

BAB 8
DESAIN SISTEM AI

8.1 TENTANG "HUKUM YANG BERLAKU" DALAM BAB INI

Garis besar resmi AIGP untuk bab ini mencakup butir berjudul "Identifikasi hukum yang berlaku untuk model AI." Meskipun ini tidak diragukan lagi merupakan pertimbangan penting dalam desain sistem AI, topik ini telah dibahas secara mendalam di bab-bab sebelumnya—khususnya di Bab 4 dan 5. Untuk menghindari pengulangan, bab ini tidak membahas kembali kerangka hukum tersebut secara detail. Pembaca dianjurkan untuk merujuk ke bab-bab tersebut untuk cakupan komprehensif tentang hukum yang berlaku, kewajiban peraturan, dan persyaratan khusus yurisdiksi. Bab ini lebih berfokus pada proses desain itu sendiri, dengan asumsi bahwa kepatuhan hukum ditangani secara paralel melalui saluran tata kelola yang tepat. Lihat bagan domain AIGP di Pendahuluan buku ini untuk lebih memahami bagaimana domain AIGP dicakup oleh bab-bab dalam buku ini.

Mendokumentasikan Siklus Hidup

Mendokumentasikan siklus hidup sistem AI sangat penting untuk memastikan kepatuhan, mengelola risiko, meningkatkan akuntabilitas, menjaga transparansi, dan memastikan kesinambungan operasional. Setiap tahap pengembangan memperkenalkan potensi risiko, kewajiban tata kelola, dan peluang untuk menanamkan praktik AI yang dapat dipercaya. Bagian ini mengeksplorasi bagaimana organisasi dapat secara sistematis mendokumentasikan setiap fase pengembangan sistem AI, dengan penekanan pada ketertelusuran, keselarasan dengan kebijakan internal dan standar eksternal, serta kesiapan untuk audit atau investigasi.

Mendokumentasikan Persyaratan

Sebelum desain sistem AI dapat dimulai, sangat penting untuk secara formal mengidentifikasi dan mencatat persyaratan sistem yang diusulkan. Dengan mendokumentasikan persyaratan, manajemen akan dapat menentukan apakah sistem beroperasi dengan benar dengan mengkonfirmasi apakah sistem tersebut berhasil memenuhi setiap persyaratan.

Organisasi dengan proses siklus hidup pengembangan sistem (SDLC) yang matang akan memiliki praktik yang mapan untuk mendokumentasikan, meninjau, dan menetapkan spesifikasi persyaratan.

Praktik penting dalam manajemen persyaratan meliputi:

- ✓ Setiap persyaratan harus dapat diverifikasi, dilacak, tidak ambigu, dan terukur.
- ✓ Kategorikan persyaratan berdasarkan jenisnya, termasuk keamanan, privasi, pengalaman pengguna (UX), peraturan, operasional, dan standar yang telah ditetapkan (OS, protokol, alat, dan sebagainya). Beberapa organisasi memiliki katalog persyaratan standar untuk diterapkan pada setiap proyek pengembangan sistem.
    - ➢ Persyaratan fungsional menjelaskan fungsi sistem AI.

- Persyaratan teknis (juga dikenal sebagai persyaratan non-fungsional) menjelaskan elemen-elemen dari tumpukan teknologi dan mencakup kinerja, skalabilitas, keamanan, privasi, kegunaan, dan ketahanan.
- Persyaratan peraturan menjelaskan fungsi dan karakteristik sistem AI yang dipersyaratkan oleh peraturan tertentu.
- Persyaratan keamanan menjelaskan fungsi dan karakteristik sistem AI yang melindungi sistem AI dan data terkait.

✓ Catat sumber, pemilik, prioritas, dan alasan untuk setiap persyaratan.
✓ Tetapkan persyaratan tata kelola data (persetujuan, dasar hukum, penanganan PII, silsilah, retensi, penghapusan).
✓ Tentukan metrik evaluasi dengan ambang batas (misalnya, akurasi, kalibrasi, latensi, biaya per inferensi).
✓ Sertakan tingkat kekritisan atau bobot setiap persyaratan.
✓ Rangkaian persyaratan akhir harus disetujui oleh manajemen.

Para perancang dan pengembang harus diberikan serangkaian persyaratan yang telah disetujui untuk dijadikan panduan selama fase perancangan dan pengembangan proyek.

Jangan Biarkan Suara Terkeras Menang

Para pemangku kepentingan dengan pengaruh organisasi dapat mendominasi proses penetapan persyaratan. Seimbangkan hal ini dengan secara eksplisit memasukkan umpan balik dari kelompok yang terpinggirkan atau terdampak. Ingatlah untuk mendokumentasikan siapa yang memberikan setiap persyaratan, beserta tingkat kepentingannya.

Mendokumentasikan Proses Perancangan dan Pembangunan

Dokumentasi untuk sistem AI memiliki dua dimensi: pertama, deskripsi rinci tentang proses perancangan dan pembangunan, dan kedua, deskripsi rinci tentang desain sistem AI. Proses perancangan dan pembangunan hanya perlu didokumentasikan sekali (dan direvisi secara berkala setelahnya), dan desain setiap sistem AI juga harus didokumentasikan. Bagian ini terutama berfokus pada dokumentasi untuk sistem AI.

Dokumentasi pada Fase Perancangan

Mendokumentasikan fase perancangan sistem AI memiliki beberapa tujuan: menciptakan catatan bisnis tentang maksud, keputusan kunci, dan membantu menyelaraskan fitur sistem dengan tujuan bisnis dan peraturan. Dalam lingkaran SDLC/tata kelola AI, ini dikenal sebagai "garis dasar desain" atau "artefak desain". Dokumentasi ini harus dimulai sejak awal konsepsi sistem dan mencatat keputusan desain saat muncul.

Elemen kunci yang perlu dicatat dalam dokumentasi desain meliputi:

❖ **Tujuan dan ruang lingkup sistem:** Mendefinisikan fungsi utama sistem AI, peran pengambilan keputusan, dan konteks penggunaan.
❖ **Pemangku kepentingan dan peran:** Identifikasi siapa yang akan terpengaruh oleh atau berinteraksi dengan sistem AI, termasuk pengguna, pelanggan, operator, pihak ketiga,

serta mereka yang bertanggung jawab untuk membangun, mengoperasikan, dan memeliharanya.

- **Pertimbangan etika dan regulasi:** Dokumentasikan bagaimana sistem selaras dengan prinsip-prinsip seperti keadilan, transparansi, akuntabilitas, privasi, ketahanan, dan kepatuhan terhadap hukum yang berlaku.
- **Batasan dan antarmuka sistem:** Klarifikasi apa yang akan dan tidak akan dilakukan sistem, bagaimana sistem akan berinteraksi dengan sistem lain, dan di mana pengawasan manusia diharapkan.
- **Desain data dan model:** Catat sumber data, kriteria kualitas, strategi pelabelan, rasional pemilihan model, kinerja target, dan batasan penjelasan.
- **Manajemen risiko dan siklus hidup:** Identifikasi risiko yang dapat diperkirakan, kasus penyalahgunaan, batasan keamanan, mekanisme cadangan, dan rencana untuk pemantauan, pelatihan ulang, dan penghentian penggunaan.

Misalnya, sistem AI yang dirancang untuk menyaring aplikasi pinjaman harus mendokumentasikan sumber datanya, bagaimana keadilan akan dipastikan dalam proses pengambilan keputusan, dan upaya hukum apa yang dimiliki pemohon untuk menantang keputusan tersebut.

Dokumentasi Minimum yang Layak untuk Desain AI

Mencakup elemen-elemen berikut akan memberikan nilai signifikan bagi proyek pengembangan:

- Pernyataan masalah (sebagaimana didefinisikan dan diungkapkan oleh bisnis, bukan oleh pengembang)
- Konteks tata kelola dan kepatuhan
- Pengguna yang dituju dan kasus penggunaan
- Asumsi dan batasan
- Input dan output sistem
- Data pelatihan yang akan digunakan
- Analisis risiko awal
- Model tingkat tinggi atau desain arsitektur

Dokumentasi dalam Fase Pengembangan

Fase pengembangan mentransisikan persyaratan dan desain ke dalam kode dan perangkat keras. Di sinilah keputusan dibuat mengenai arsitektur sistem, alur kerja pengumpulan data, rekayasa fitur, dan pemilihan model awal. Pada fase ini, organisasi harus mengembangkan:

- **Kartu model dan kartu sistem:** Dibahas dalam bab-bab sebelumnya, ini adalah artefak khusus AI yang penting yang merinci karakteristik model AI, kasus penggunaan yang dimaksudkan, keterbatasan, pertimbangan etis, dan kinerja.
- **Diagram arsitektur:** Sertakan visual tingkat tinggi dan detail dari komponen sistem dan interaksinya, dengan kontrol versi.

- **Repositori kode dan model:** Sertakan keamanan yang tepat, kontrol sumber, termasuk cabang, riwayat perubahan, dan catatan kontributor.
- **Perangkat dan dependensi:** Lacak pustaka perangkat lunak, layanan cloud, dan kerangka kerja yang digunakan, yang semuanya penting untuk debugging, patching, dan upaya pemeliharaan dan pengembangan di masa mendatang.
- **Pertimbangan keamanan:** Pastikan kode dan data terlindungi selama pengembangan, termasuk penggunaan sandboxing atau secure enclave dan pemindaian kerentanan. Semua artefak ini harus berada di repositori catatan bisnis resmi dan tunduk pada tinjauan dan persetujuan formal.

Contoh tim membangun chatbot pemrosesan bahasa alami (NLP). Tim tersebut mendokumentasikan langkah-langkah pra-pemrosesan yang diterapkan pada kueri pengguna, metode tokenisasi yang digunakan, dan model bahasa milik perusahaan yang disempurnakan.

Mendokumentasikan Proses Pelatihan dan Pengujian

Setelah arsitektur sistem AI diselesaikan dan disetujui, fase selanjutnya melibatkan pelatihan, validasi, dan pengujian model, yang merupakan tahapan penting untuk menyempurnakan kinerja model, mendeteksi bias, dan memvalidasi output.

Dokumentasi lengkap pada tahap ini membantu memastikan bahwa hasil pengujian dapat diulang, kewajiban kebijakan dan peraturan dipenuhi, dan perilaku yang tidak terduga dapat diidentifikasi dan diselesaikan.

Dokumentasi Pelatihan

Dokumentasi pelatihan harus mencakup detail tentang bagaimana model dilatih, serta informasi latar belakang, termasuk alasan di balik keputusan pelatihan tertentu dan bagaimana pertimbangan kinerja ditangani. Dokumentasi pelatihan harus mencakup:

- **Sumber dan karakteristik data pelatihan:** Sertakan semua sumber data, pengecualian, penambahan, penyaringan, dan populasi yang diwakili. Semua data pelatihan harus diarsipkan dan tidak dapat diubah untuk memastikan pelatihan ulang atau pemecahan masalah di masa mendatang dimungkinkan.
- **Proses pelabelan dan jaminan kualitas:** Jika pembelajaran terawasi terlibat, catat bagaimana label diberikan dan diverifikasi.
- **Konfigurasi pelatihan model:** Sertakan hyperparameter, ukuran batch, jumlah epoch, strategi optimasi, dan metrik serta hasil evaluasi.
- **Tolok ukur kinerja model:** Kinerja dasar, metrik berkelanjutan, dan informasi yang mendukung ambang batas yang dipilih.
- **Deteksi dan mitigasi bias:** Catat penilaian keadilan, teknik de-biasing, dan trade-off.
- **Hasil pengujian:** Catat hasil pengujian ketahanan, generalisasi, dan privasi/keamanan untuk memastikan model berkinerja andal dalam kondisi yang diharapkan dan tidak diharapkan.

Misalnya, tim yang melatih model visi komputer untuk mendeteksi cacat pada komponen yang diproduksi membuat dokumentasi yang menjelaskan bagaimana 20.000 gambar diperoleh dari berbagai jalur produksi, diberi label oleh teknisi terlatih, dan diperkaya menggunakan filter rotasi dan kontras.

Harapan Regulasi untuk Log Pelatihan

Dalam sistem AI berisiko tinggi berdasarkan Undang-Undang AI Uni Eropa dan kerangka kerja lainnya, dokumentasi pelatihan harus mendukung:

- Dokumentasi komprehensif yang cukup untuk mereproduksi upaya desain dan pengembangan
- Penjelasan tentang rasionalitas desain
- Kemampuan audit jangka panjang
- Bukti mitigasi risiko

Dokumentasi Pengujian dan Validasi

Pengujian mengungkapkan keandalan dan keamanan suatu model. Seperti pelatihan, pengujian harus didokumentasikan dengan baik, untuk memverifikasi kinerja dan memberikan bukti manajemen risiko. Dokumentasi pengujian harus mencakup hal-hal berikut:

- **Metodologi pengujian:** Cantumkan jenis pengujian yang dijalankan (unit, integrasi, kinerja, adversarial, dan sebagainya). NIST ARIA, yang dibahas dalam Bab 7, mencakup panduan tentang pengujian model.
- **Deskripsi data uji:** Nyatakan apakah data uji ditahan dari set pelatihan dan apakah data tersebut mewakili kondisi dunia nyata.
- **Hasil pengujian:** Sertakan akurasi, positif/negatif palsu, presisi/recall, dan metrik lainnya.
- **Pengujian stres dan analisis kasus ekstrem:** Catat kondisi batas, contoh adversarial, dan input serta output di luar norma yang diharapkan.
- **Mode kegagalan dan respons:** Sertakan kelemahan yang diketahui dan detail perilaku sistem jika terjadi kegagalan.
- **Pengujian pengguna dan umpan balik:** Sertakan detail tentang bagaimana kegunaan dan interaksi manusia-AI dievaluasi.

Misalnya, sebuah organisasi yang mengembangkan alat perekrutan berbasis AI menguji sistem sebelum penerapan menggunakan dataset sintetis untuk mengeksplorasi bagaimana sistem menangani kasus ekstrem, seperti resume yang tidak lengkap, format yang tidak konsisten, atau riwayat pekerjaan yang tidak biasa.

Pengujian Kepercayaan

AI yang dapat dipercaya membutuhkan lebih dari sekadar akurasi tinggi. Pengujian harus menjawab pertanyaan-pertanyaan berikut:

- ✓ Apakah sistem gagal dengan baik?
- ✓ Dapatkah output dijelaskan atau dibenarkan?
- ✓ Apakah pola kesalahan dapat diterima berdasarkan standar etika dan hukum?

Mempertahankan Keterlacakan dan Kesiapan Audit

Dokumentasi menyeluruh atas pengembangan sistem apa pun sangat penting, dan bisa dibilang sangat penting untuk sistem AI, mengingat serangkaian risiko unik yang terkait dengan AI. Dokumentasi komprehensif memungkinkan organisasi untuk melacak perilaku AI kembali ke asalnya, kemampuan penting saat menanggapi audit, pertanyaan publik, atau litigasi. Keterlacakan membangun kepercayaan, mendukung kepatuhan, dan memfasilitasi manajemen risiko jangka panjang yang lebih efektif.

Kontrol Versi dan Manajemen Perubahan

Semua perubahan dalam persyaratan, desain, data, dan konfigurasi model harus dilacak dengan cermat. Praktik penting:

- Mempertahankan kontrol versi yang terhubung di seluruh kode, data, dan dokumentasi. Membutuhkan tinjauan sejawat dan persetujuan formal untuk perubahan utama.
- Memastikan persyaratan terhubung dengan elemen desain, rencana pengujian, dan hasil pengujian.
- Mencatat tidak hanya keputusan tetapi juga alasan dan hasilnya.
- Memberi tag pada rilis dengan penilaian risiko atau persetujuan terkait.

Kemampuan Menjelaskan dan Artefak Dokumentasi

Sebagai fitur model, kemampuan menjelaskan memiliki implikasi dokumentasi. Organisasi perlu membuat dan menyimpan artefak yang membantu mereka menafsirkan bagaimana model AI membuat keputusan. Artefak penting meliputi:

- Alur atau diagram pengambilan keputusan yang dianotasi
- Ringkasan pentingnya fitur
- Justifikasi atau pembalikan tingkat kasus
- Jalur dan hasil eskalasi Human-in-the-loop (HITL)
- Kartu model dan sistem

Integrasi dengan Kerangka Tata Kelola yang Lebih Luas

Seperti yang disebutkan di seluruh buku ini, tata kelola AI harus diintegrasikan dengan tata kelola pendukung lainnya, termasuk TI, keamanan siber, dan privasi. Dokumentasi sistem AI harus terhubung ke sistem tata kelola organisasi yang lebih luas, termasuk:

- Manajemen risiko
- Inventaris model
- Respons insiden
- Proses dewan peninjauan etika

Menyelaraskan dokumentasi dengan kontrol internal membantu membuat tata kelola lebih mudah diulang, diaudit, dan dipertahankan. Ketika praktik dokumentasi diintegrasikan dengan sistem risiko dan kepatuhan perusahaan, praktik tersebut tidak hanya mendukung kesiapan regulasi tetapi juga menciptakan fondasi yang dapat diskalakan untuk pengembangan AI di masa mendatang.

8.2 KONTEKS BISNIS DAN KASUS PENGGUNAAN

Efektivitas sistem AI bergantung pada kinerja teknisnya dan apakah sistem tersebut secara efektif melayani tujuan bisnis yang dimaksudkan. Menentukan konteks bisnis dan kasus penggunaan spesifik untuk model AI merupakan langkah mendasar dalam pengembangan yang bertanggung jawab, dan harus diselesaikan secara formal sebelum menyelesaikan persyaratan dan desain, serta sebelum pengembangan. Konteks dan kasus penggunaan menjadi landasan tujuan organisasi sistem, memperjelas harapan kinerja, dan menyoroti kendala operasional, peraturan, dan etika. Tanpa konteks bisnis dan kasus penggunaan yang terdokumentasi, bagaimana sebuah organisasi dapat diharapkan untuk mengembangkan sistem AI yang sesuai? Bagian ini membahas konteks bisnis dan menguraikan kasus penggunaan, serta cara mengevaluasi kesiapan dan kelayakan yang lebih luas untuk menerapkan AI di lingkungan dunia nyata.

Menentukan Konteks Bisnis

Konteks bisnis adalah elemen mendasar dari pengembangan sistem AI, karena mendefinisikan lingkungan organisasi, strategis, dan operasional di mana sistem AI akan digunakan. Demikian pula, kasus bisnis adalah pernyataan tentang manfaat yang diharapkan bagi organisasi yang diantisipasi akan terwujud sebagai hasil dari pengembangan sistem AI.

Memahami Tujuan Organisasi

Setiap proyek AI harus dimulai dengan tujuan bisnis yang diartikulasikan dengan jelas—alasan keberadaan sistem AI dan kontribusi yang diharapkan bagi organisasi. Hal ini membantu memastikan bahwa tim teknis tidak menciptakan sistem yang mencari masalah, tetapi membangun solusi yang selaras dengan kebutuhan strategis.

Pertimbangan utama dalam mendukung konteks bisnis meliputi:

- Masalah yang ingin dipecahkan oleh sistem AI
- Metrik keberhasilan yang diharapkan dari perspektif bisnis
- Bagaimana sistem akan mendukung tujuan organisasi yang lebih luas

Contoh perusahaan asuransi berupaya meningkatkan deteksi penipuan dan bermaksud menerapkan sistem pembelajaran mesin untuk menandai klaim yang mencurigakan. Tujuan bisnisnya adalah untuk mengidentifikasi data yang menyimpang dan mengurangi kerugian akibat penipuan dengan persentase yang terukur, sambil meminimalkan kesalahan positif.

Kesiapan dan Tata Kelola Organisasi

Sebelum memulai pengembangan, organisasi harus menilai kemampuan mereka untuk mendukung dan mengoperasikan sistem AI, termasuk infrastruktur, kebijakan, keselarasan kepemimpinan, dan proses bisnis. Pertimbangan meliputi:

- Dukungan kepemimpinan dan komitmen anggaran berkelanjutan
- Kematangan tata kelola data
- Kematangan keamanan siber dan privasi
- Toleransi risiko dan budaya etika
- Ketersediaan keahlian lintas fungsi (hukum, TI, ilmu data, operasi, keamanan siber, dan privasi)

Penilaian Risiko Proyek AI

Beberapa organisasi menetapkan tingkatan risiko untuk proyek AI selama perencanaan. Model yang berdampak tinggi dan berhadapan langsung dengan pelanggan mungkin memerlukan dokumentasi yang lebih kuat, tinjauan pra-implementasi, dan dukungan eksekutif daripada alat internal berisiko rendah. Tingkatan risiko yang didefinisikan dalam Undang-Undang AI Uni Eropa (dibahas dalam Bab 6) adalah permulaan yang baik, bahkan untuk organisasi yang tidak tunduk pada undang-undang tersebut (peraturan di masa mendatang mungkin menggunakan Undang-Undang AI Uni Eropa sebagai model).

Identifikasi dan Keterlibatan Pemangku Kepentingan

Memahami siapa yang akan berinteraksi dengan, mendapat manfaat dari, atau terpengaruh oleh sistem AI sangat penting untuk membentuk ruang lingkup, persyaratan, batasan, dan metode evaluasinya. Pemangku kepentingan umumnya meliputi:

- Sponsor bisnis (biasanya, eksekutif senior)
- Pengembang sistem dan ilmuwan data
- Tim hukum dan kepatuhan
- Tim keamanan siber dan privasi
- Tata kelola data
- Regulator dan auditor eksternal
- Pengguna akhir dan pelanggan
- Individu yang secara tidak langsung terdampak oleh keputusan (misalnya, pelamar yang ditolak)

Misalnya, untuk model AI penerimaan universitas, pemangku kepentingan yang diidentifikasi meliputi petugas penerimaan (pengguna), pelamar (subjek), regulator (pengawasan), dan kepemimpinan institusional (sponsor).

Mengartikulasikan Kasus Penggunaan AI

Kasus penggunaan yang terdefinisi dengan baik menerjemahkan konteks bisnis ke dalam spesifikasi tingkat tinggi untuk aplikasi AI. Kasus penggunaan menjelaskan bagaimana sistem akan beroperasi, data apa yang akan digunakan, dan hasil apa yang diharapkan untuk diberikan. Hal ini memungkinkan pemangku kepentingan untuk menilai kelayakan, persyaratan sumber daya, persyaratan kepatuhan, dan implikasi etis sebelum berkomitmen pada desain dan pengembangan.

Menentukan Lingkup dan Batasan

Sangat penting bagi organisasi untuk mendefinisikan apa yang akan dilakukan sistem AI, dan apa yang tidak akan dilakukannya. Definisi yang jelas membantu menghindari perluasan lingkup, ekspektasi yang tidak sesuai, dan risiko yang tidak terkendali. Definisi lingkup harus mencakup:

- ✓ Fungsi yang dimaksud (misalnya, klasifikasi, prediksi, bantuan, rekomendasi)
- ✓ Pengguna dan lingkungan yang dimaksud
- ✓ Asumsi dan batasan (misalnya, sumber data, infrastruktur, batasan peraturan)

- ✓ Otoritas pengambilan keputusan (apakah sistem bertindak secara otonom, atau apakah ada campur tangan manusia)
- ✓ Masa pakai sistem yang diharapkan, beban kerja, dan frekuensi penggunaan

Contoh pengecer yang berencana menggunakan AI untuk penetapan harga dinamis harus menentukan apakah keputusan penetapan harga hanya akan diterapkan secara online atau di toko, apakah harga akan diperbarui setiap jam atau setiap hari, dan apakah diperlukan tinjauan manusia untuk penyimpangan yang signifikan.

Menentukan Metrik Keberhasilan

Kurangnya metrik yang jelas menimbulkan risiko bahwa organisasi akan mengoptimalkan sistem AI untuk hasil yang salah. Metrik harus selaras dengan tujuan bisnis, realistis berdasarkan data yang tersedia, dan peka terhadap konsekuensi yang tidak diinginkan. Contoh metrik dapat mencakup:

- ❖ **Untuk deteksi penipuan:** Pengurangan penipuan yang tidak terdeteksi, tingkat positif palsu
- ❖ **Untuk efisiensi chatbot:** Waktu penyelesaian, tingkat eskalasi, kepuasan pelanggan
- ❖ **Untuk alat perekrutan:** Kualitas perekrutan, waktu pengisian posisi, bias di seluruh kelas yang dilindungi

Jebakan Metrik

Terlalu bergantung pada presisi dapat menyebabkan peningkatan kesalahan negatif. Mengoptimalkan efisiensi dapat mengurangi transparansi. Selalu pertimbangkan konsekuensinya.

Menilai Ketersediaan dan Kesesuaian Data

Kegagalan banyak proyek AI bukan karena model atau desain yang cacat, melainkan karena ketersediaan data yang buruk atau ketidaksesuaian antara data dan tujuan bisnis. Mengevaluasi kesesuaian data di awal proyek memastikan kasus penggunaan layak dan risiko dapat dikelola. Kesesuaian data harus diselesaikan sebelum upaya pengembangan dimulai.

Menginventarisasi Data yang Tersedia

Organisasi harus menilai data apa yang sudah mereka miliki, di mana data tersebut berada, dan bagaimana data tersebut dikumpulkan. Lebih lanjut, data yang menjadi kandidat harus dievaluasi kelengkapannya, keakuratannya, konsistensinya, dan ketersediaan labelnya jika data tersebut akan digunakan untuk pembelajaran terawasi. Misalnya, sebuah bank yang ingin memprediksi gagal bayar pinjaman mungkin menemukan bahwa data pendapatan pelanggan dikumpulkan secara tidak konsisten, yang memerlukan pembersihan data tambahan atau revisi asumsi model.

Tata Kelola Data Merupakan Prasyarat Utama

Organisasi yang tidak memiliki kebijakan, praktik, dan alat tata kelola data yang matang akan berada dalam posisi yang sangat tidak menguntungkan, karena mereka hanya memiliki kesadaran parsial tentang inventaris data dan proses manajemen mereka.

Batasan Hukum dan Etika dalam Penggunaan Data

Meskipun data pelatihan yang sesuai tersedia, mungkin ada alasan hukum atau etika mengapa data tersebut tidak dapat digunakan apa adanya. Kekhawatiran utama dapat meliputi:

- Data yang dikumpulkan untuk satu tujuan digunakan kembali oleh sistem AI tanpa persetujuan subjek
- Penyertaan atribut sensitif (misalnya, data ras, kesehatan, atau afiliasi agama)
- Kesimpulan tentang individu yang dapat melanggar harapan privasi

Dampak Bisnis dari Pasal 5(1)(b) GDPR

"Pembatasan tujuan" dalam GDPR dapat mencegah organisasi menggunakan data historis untuk pelatihan, kecuali jika dasar hukum baru untuk pemrosesan dapat ditetapkan.

Kesiapan Data untuk Penerapan

Organisasi perlu menilai apakah ekosistem data dapat mendukung kebutuhan data yang berkelanjutan, tidak hanya untuk pelatihan tetapi juga untuk pemantauan, pelatihan ulang, dan kepatuhan. Pertimbangan meliputi:

- Apakah alur data stabil dan dapat diskalakan
- Apakah model akan diberi data baru dan terbaru dari waktu ke waktu
- Apakah inferensi waktu nyata dapat didukung, jika diperlukan
- Apakah pemantauan telah dilakukan untuk mendeteksi pergeseran atau degradasi data
- Apakah penyerapan data berkelanjutan sesuai dengan persyaratan privasi, keamanan, dan kontrol akses

Pertimbangan Etika dan Operasional

Sangat mungkin bahwa kasus penggunaan AI yang layak dan berharga masih akan menimbulkan tanda bahaya etika atau operasional. Pertimbangan penting ini membantu organisasi memastikan bahwa penerapannya bertanggung jawab, berkelanjutan, dan selaras dengan nilai-nilai organisasi.

Risiko Kerugian dan Pertukaran Etika

Meskipun legal, beberapa kasus penggunaan dapat menyebabkan hasil yang merugikan individu, memperburuk ketidaksetaraan, atau mengurangi kemampuan bertindak, yang semuanya dapat membuat pihak yang terkena dampak marah dan menarik perhatian regulator. Area-area yang menjadi perhatian umum meliputi:

- Penggunaan AI dalam peradilan pidana, perumahan, perekrutan, pendidikan, atau kredit
- Penguatan bias historis
- Ketidaktransparansian dalam keputusan-keputusan penting
- Manipulasi perilaku secara algoritmik (misalnya, lingkaran kecanduan di media sosial)

Misalnya, chatbot kesehatan mental bertenaga AI mungkin mengurangi biaya tetapi gagal untuk meningkatkan penanganan pengguna yang sedang krisis secara tepat, sehingga

menciptakan risiko serius. Protokol eskalasi mungkin sangat penting dalam kasus penggunaan seperti ini.

Pengawasan Manusia dan Hak Pengambilan Keputusan

Setiap sistem AI yang membuat keputusan bisnis harus dirancang sebagai bagian dari proses pengambilan keputusan yang berpusat pada manusia. Pertimbangkan hal-hal berikut:

- Keputusan apa yang dapat dibuat AI secara otonom?
- Kapan manusia harus terlibat dalam pengambilan keputusan?
- Bagaimana penanganan pembatalan atau banding?
- Apakah peran pengawasan manusia didefinisikan dengan baik dan didukung sumber daya yang memadai?
- Apakah ada catatan bisnis yang cukup untuk membela keputusan dan tindakan lainnya?

Matriks RACI untuk Pengawasan AI

Berikut adalah peran yang disarankan untuk pengawasan manusia terhadap sistem AI:

- Bertanggung jawab: Pemimpin ilmu data
- Akuntabel: Pemilik produk tingkat C
- Dikonsultasikan: Komite hukum dan etika, petugas privasi, petugas keamanan, dan manajer risiko
- Diberi informasi: Manajer lini depan

Kesiapan dan Budaya Kerja

Ketika sistem AI diperkenalkan ke dalam organisasi, sistem tersebut tidak hanya mengubah alur kerja; tetapi juga mengubah peran, persyaratan pekerjaan, dan budaya organisasi. Poin-poin yang perlu dipertimbangkan terkait kesiapan dan budaya organisasi meliputi:

- ✓ Apakah tim yang terdampak telah dikonsultasikan dan masukan mereka telah didengar;
- ✓ Apakah pelatihan ulang atau perekrutan baru diperlukan;
- ✓ Apakah alat bantu kerja, sistem pendukung, dan jalur eskalasi telah tersedia;
- ✓ Apakah keterlibatan karyawan secara teratur menunjukkan perubahan moral;

Keselarasan dengan Nilai dan Strategi Organisasi

Kasus penggunaan AI mungkin secara teknis sangat baik dan secara hukum sah. Namun, hal itu mungkin tidak selaras dengan identitas merek, reputasi, atau komitmen publik organisasi. Hal-hal yang perlu dipertimbangkan meliputi:

- ❖ Apakah sistem AI atau kasus penggunaan tertentu bertentangan dengan pernyataan publik atau tujuan perusahaan?
- ❖ Apakah sistem AI akan mengikis kepercayaan pelanggan atau mitra?
- ❖ Apakah sistem AI dapat menimbulkan risiko reputasi yang tidak proporsional dengan nilainya?
- ❖ Apakah staf akan melihat proyek tersebut sesuai dengan nilai-nilai perusahaan atau dapat memicu penolakan internal?

8.3 PENILAIAN DAMPAK

Penilaian dampak merupakan landasan tata kelola AI yang bertanggung jawab. Baik diwajibkan oleh peraturan atau didorong oleh kebijakan internal atau hasil manajemen risiko, penilaian dampak membantu organisasi mengantisipasi, mengukur, dan mengurangi potensi konsekuensi dari penerapan sistem AI sebelum penerapannya.

Melakukan atau meninjau penilaian dampak bukanlah aktivitas sekali saja, tetapi merupakan proses terstruktur dan berulang yang menyelaraskan sistem AI dengan prinsip-prinsip etika, persyaratan hukum dan peraturan, serta harapan pemangku kepentingan.

Memahami Penilaian Dampak AI

Tujuan utama dari penilaian dampak AI (AIIA) adalah untuk mengidentifikasi dan mengelola potensi kerugian atau konsekuensi yang tidak diinginkan sebelum penerapan. Ini dapat mencakup diskriminasi, kurangnya transparansi, kegagalan keselamatan, pengawasan berlebihan, atau pelanggaran hak-hak fundamental.

Ketika didokumentasikan sebagai proses yang dapat diulang, penilaian dampak memberikan metode yang konsisten untuk identifikasi risiko dan mendorong transparansi serta akuntabilitas. Ketika didokumentasikan, penilaian dampak berfungsi sebagai dokumentasi untuk tata kelola internal dan kepatuhan peraturan. Hal ini dapat membantu meningkatkan kepercayaan pemangku kepentingan dan penerimaan publik. Misalnya, sebuah lembaga sektor publik yang menerapkan teknologi pengenalan wajah bertenaga AI menggunakan penilaian dampak untuk mengeksplorasi risiko privasi, potensi bias, dan kebutuhan konsultasi publik. Hasil penilaian tersebut mengarah pada implementasi perlindungan baru sebelum penerapan.

Pendorong Hukum dan Regulasi

Banyak yurisdiksi sekarang mewajibkan atau merekomendasikan penilaian dampak untuk sistem AI, terutama yang dianggap berisiko tinggi. Contohnya meliputi:

- **Undang-Undang AI Uni Eropa:** Pasal 6 mewajibkan "penilaian kesesuaian" dan dokumentasi untuk sistem AI berisiko tinggi.
- **Arahan Kanada tentang Pengambilan Keputusan Otomatis (DADM):** Pasal 6.1 mewajibkan Penilaian Dampak Algoritma (AIA) untuk sistem pengambilan keputusan berbasis AI.
- **CPRA California (dan undang-undang serupa):** Memberikan wewenang kepada regulator untuk mewajibkan penilaian risiko untuk pengambilan keputusan otomatis.
- **Prinsip-prinsip OECD dan G7:** Mendorong penilaian dampak sebagai praktik terbaik.

Tidak Semua Penilaian Risiko Sama

Meskipun penilaian dampak keamanan siber, perlindungan data, dan AI mungkin tumpang tindih, masing-masing berfokus pada domain risiko yang berbeda. Menyelaraskannya dapat menyederhanakan tata kelola dan mengurangi redundansi.

Kapan dan Seberapa Sering Melakukan Penilaian

Penilaian dampak AI harus dilakukan selama fase desain sistem AI sebelum penerapan dan setiap kali perubahan signifikan dilakukan pada model atau datanya, serta sebagai respons terhadap insiden. Sistem AI berisiko tinggi dan berdampak tinggi harus dinilai secara teratur.

Persiapan untuk Penilaian Dampak

Organisasi perlu mempersiapkan penilaian dampak yang akan datang. Pertimbangan utama meliputi mendefinisikan ruang lingkup, mengidentifikasi pemangku kepentingan, dan menentukan informasi yang diperlukan. Hal-hal ini harus didokumentasikan sebelum dimulainya penilaian.

Mendefinisikan Ruang Lingkup Penilaian

Ruang lingkup penilaian memastikan penilaian tetap fokus dan selaras dengan tujuan penggunaan sistem AI. Definisi ruang lingkup harus mencakup:

- Deskripsi dan fungsionalitas sistem
- Konteks bisnis
- Kasus penggunaan (untuk sistem AI dengan lebih dari satu)
- Pengguna dan populasi yang terpengaruh
- Yurisdiksi geografis dan hukum
- Tingkat otonomi dan pengawasan manusia dalam proses bisnis yang didukung
- Asumsi, batasan, dan pengecualian

Misalnya, model AI yang digunakan untuk tinjauan kinerja karyawan mungkin perlu dinilai secara berbeda di Uni Eropa (karena Pasal 22 GDPR tentang keputusan otomatis yang dibuat semata-mata oleh AI) daripada di Amerika Serikat, di mana hukum ketenagakerjaan berbeda.

Identifikasi Pemangku Kepentingan

Pemangku kepentingan bisnis dan teknis harus dilibatkan dalam tim penilaian untuk memastikan bahwa semua perspektif terwakili. Peserta penilaian dapat mencakup:

- Sponsor proyek dan pemilik produk
- Pengembang dan ilmuwan data
- Petugas hukum, keamanan, kepatuhan, dan privasi
- Pakar domain (misalnya, etika, DEI, aksesibilitas)
- Perwakilan kelompok yang terdampak, jika memungkinkan

Beberapa organisasi mungkin ingin menyertakan pemangku kepentingan eksternal, seperti penasihat, konsultan, vendor, dan anggota masyarakat.

Inventarisasi Informasi yang Diperlukan

Sebelum penilaian dampak dapat dilakukan, tim harus mengumpulkan informasi dasar, termasuk:

- Dokumentasi organisasi (kebijakan, proses, kode etik)
- Dokumentasi sistem (desain, fungsionalitas, persyaratan, tujuan penggunaan)
- Hukum, peraturan, dan kewajiban hukum lainnya yang berlaku
- Pemangku kepentingan (identifikasi pengguna, operator, dan pihak yang terdampak)
- Alur data, sumber, asal usul, dan jenisnya
- Metrik kinerja model dan keterbatasan yang diketahui

- Detail pengalaman pengguna
- Insiden sebelumnya atau kekhawatiran publik di domain serupa

Melakukan Penilaian Dampak

Proses penilaian dampak itu sendiri merupakan serangkaian langkah: mengidentifikasi potensi kerugian, mengevaluasi risiko, mendokumentasikan mitigasi, dan menangkap kekhawatiran yang tersisa. Bagian ini menguraikan langkah-langkah dan alat penilaian umum. Proses untuk melakukan AIA harus didokumentasikan, termasuk mengidentifikasi pihak yang bertanggung jawab, dan lokasi tempat hasil akan disimpan.

Mengidentifikasi Potensi Dampak

AIA dimulai dengan memetakan di mana dan bagaimana sistem AI dapat menimbulkan risiko atau kerugian. Area-area yang berpotensi terkena dampak meliputi:

- Privasi dan pengawasan
- Pencurian kekayaan intelektual atau data privasi
- Diskriminasi dan bias
- Kebebasan berekspresi atau berasosiasi
- Otonomi dan martabat manusia
- Kerugian atau manipulasi psikologis
- Akses terhadap hak atau kesempatan
- Dampak lingkungan

Misalnya, alat penyaringan pekerjaan berbasis AI dapat menghukum pelamar dengan celah dalam riwayat pekerjaan mereka, secara tidak langsung menghukum pengasuh, penyandang disabilitas, atau veteran.

Evaluasi Kemungkinan dan Tingkat Keparahan Risiko

Untuk setiap dampak yang diidentifikasi, penting untuk menilai:

- ✓ Kemungkinan: Probabilitas dampak
- ✓ Tingkat Keparahan: Tingkat kerugian bagi individu atau kelompok
- ✓ Cakupan: Jumlah orang yang terpengaruh
- ✓ Kemungkinan Pemulihan: Apakah kerugian dapat diperbaiki

Alat penilaian, seperti matriks risiko atau peta panas, dapat membantu memprioritaskan risiko yang diidentifikasi.

Analisis Risiko Diskriminasi dan Keadilan

Karena sifat model AI, risiko bias dan keadilan memerlukan perhatian khusus. Sistem AI dapat melakukan diskriminasi, bahkan ketika atribut yang dilindungi, seperti ras atau jenis kelamin, dikecualikan dari data pelatihan. Pertimbangan penilaian meliputi:

- ❖ Apakah sistem tersebut secara tidak proporsional memengaruhi kelompok yang dilindungi;
- ❖ Apakah metrik keadilan dianalisis;
- ❖ Apakah variabel proksi atau bias historis tertanam dalam data pelatihan;

Bahkan data input netral pun dapat menghasilkan output diskriminatif jika sistem AI dilatih berdasarkan pola historis yang bias. Sangat penting untuk menilai hasil, bukan hanya fitur.

Menilai Kemampuan Penjelasan dan Transparansi

Dalam bisnis, seringkali sangat penting bagi organisasi untuk dapat memahami dan menjelaskan bagaimana sistem AI melakukan apa yang dilakukannya. Jika pemangku kepentingan tidak dapat memahami bagaimana atau mengapa sistem AI membuat keputusan tertentu, organisasi mungkin akan mengalami masalah kepercayaan dan akuntabilitas. Berikut beberapa pertimbangan penting:

- Apakah pengguna akhir dapat memahami cara kerja sistem?
- Apakah keputusan dapat dijelaskan kepada individu yang terkena dampak?
- Apakah dokumentasi tersedia untuk regulator atau auditor?

Misalnya, koperasi kredit yang ingin menerapkan sistem evaluasi aplikasi pinjaman berbasis AI harus menyertakan persyaratan mengenai kemampuan penjelasan, beserta dokumentasi dan catatan untuk menunjukkan hal ini.

Meninjau Mekanisme Pengawasan Manusia

Sistem AI tidak boleh diluncurkan tanpa pengawasan manusia yang berkelanjutan, termasuk kemampuan manusia dalam proses (human-in-the-loop) untuk tindakan dan keputusan berbasis AI. Sangat penting untuk dapat menentukan apakah sistem tersebut menyediakan intervensi manusia yang efektif. Beberapa pertanyaan yang perlu diajukan selama penilaian meliputi:

- Apakah manusia "terlibat" atau "tidak terlibat"?
- Bisakah keputusan ditinjau, diajukan banding, atau dibatalkan?
- Apakah personel dilatih dan diberi wewenang untuk mengesampingkan hasil?
- Apakah pengesampingan keputusan dipandu oleh kebijakan dan didokumentasikan?

Pengawasan Bukan Sekadar Centang

Dalam proses bisnis berbasis AI yang terkelola, peninjau manusia harus memiliki waktu untuk turun tangan, wewenang untuk membatalkan keputusan, dan visibilitas tentang bagaimana keputusan dibuat. Pengawasan, baik itu melibatkan manusia dalam proses atau peninjauan pasca-keputusan, harus menjadi proses bisnis standar, bukan aktivitas sesekali atau yang didorong oleh insiden.

Strategi Mitigasi Dokumen

Ketika AIA mengidentifikasi risiko, setiap risiko yang teridentifikasi harus didokumentasikan, dianalisis, dan disertai dengan rencana mitigasi. Contoh mitigasi risiko meliputi:

- Penghapusan fitur sensitif atau variabel proksi
- Penerapan algoritma yang memperhatikan keadilan, penghapusan bias yang merugikan, atau teknik penjelasan
- Penambahan mekanisme persetujuan pengguna atau penolakan
- Audit rutin atau pengujian bias
- Peningkatan keputusan berisiko tinggi kepada peninjau manusia

Seperti manajemen risiko pada umumnya, beberapa risiko dapat diterima, dan semua contoh penerimaan risiko harus didokumentasikan, termasuk pihak yang menyetujui risiko tersebut. Risiko residual juga harus didokumentasikan, bersama dengan setiap kompromi yang diterima.

Meninjau dan Mengoperasionalkan Penilaian

Seperti penilaian risiko lainnya, penilaian dampak hanya berharga jika temuannya ditinjau, ditindaklanjuti, dan diintegrasikan ke dalam tata kelola AI dan risiko yang lebih luas. Bagian terakhir ini menjelaskan cara memastikan penilaian bermakna dan dapat ditindaklanjuti.

Tinjauan dan Persetujuan Internal

Setelah selesai, penilaian harus ditinjau oleh individu atau komite dengan wewenang tata kelola (wewenang tersebut harus didokumentasikan dalam dokumentasi tata kelola AI). Aktivitas selama peninjauan dapat mencakup:

- Pemeriksaan kepatuhan hukum dan peraturan
- Proses penerimaan atau eskalasi risiko
- Tinjauan etika (jika berlaku)
- Pengesahan tingkat eksekutif atau dewan untuk sistem berisiko tinggi

Di beberapa organisasi, peninjauan dan persetujuan AIA untuk sistem AI berisiko tinggi dapat ditugaskan ke tingkat kepemimpinan organisasi yang lebih tinggi daripada untuk sistem berisiko rendah. Pengelompokan risiko sistem AI harus diperluas ke AIA (Analisis Dampak AI), dengan mempertimbangkan seberapa sering dan ketat AIA dilakukan, serta tingkat manajemen yang diperlukan untuk meninjau dan menyetujuinya.

Integrasi dengan Pengembangan dan Penerapan

Temuan dari penilaian dampak harus dimasukkan ke dalam proses pengembangan dan penerapan. Di sinilah siklus umpan balik AIA mengarah pada perubahan aktual dalam sistem AI. Aktivitas yang mungkin dihasilkan dari AIA meliputi:

- Perubahan pada fitur model atau data pelatihan
- Penyesuaian pada antarmuka pengguna untuk penjelasan yang lebih baik
- Revisi jadwal penerapan atau strategi peluncuran
- Penambahan mekanisme pemantauan atau peringatan baru

Misalnya, penerapan sistem moderasi konten AI ditunda setelah penilaian dampak mengidentifikasi risiko sensor berlebihan dalam bahasa minoritas, yang mengakibatkan upaya pelatihan ulang untuk model tersebut.

Transparansi dan Akuntabilitas Eksternal

Tergantung pada konteks peraturan dan kewajiban kontraktual, organisasi mungkin diharuskan untuk mempublikasikan hasil atau ringkasan penilaian dampak mereka. Metode yang tersedia meliputi:

- ✓ Ringkasan publik (seperti yang terlihat pada Penilaian Dampak Algoritma Kanada)
- ✓ Pengajuan kepada badan pengatur (misalnya, berdasarkan Undang-Undang AI Uni Eropa)
- ✓ Pengarahan pemangku kepentingan atau konsultasi komunitas

Teknik-teknik ini sering digunakan untuk bentuk audit dan penilaian lainnya, termasuk uji penetrasi, penilaian risiko, Sarbanes-Oxley, dan audit PCI DSS.

Apa yang Tidak Boleh Dipublikasikan

Organisasi dilarang mempublikasikan arsitektur model yang bersifat rahasia, data yang dilindungi, atau detail yang berlebihan dalam AIA yang dapat dieksploitasi oleh pihak lain jika diketahui.

Penilaian Ulang Berkelanjutan

Karena sistem AI berubah seiring waktu, penilaian dampak harus dianggap sebagai aktivitas periodik standar, dan bahkan sebagai pemantauan berkelanjutan. AIA harus ditinjau kembali ketika:

- Model AI dilatih ulang atau diterapkan kembali
- Sumber data baru ditambahkan
- Teknik penyaringan data diubah
- Kasus penggunaan atau lingkungan berubah
- Risiko baru muncul di lapangan
- Peraturan baru diberlakukan

Misalnya, sistem pemantauan lalu lintas bertenaga AI yang awalnya dilatih pada data cuaca kering harus dinilai ulang setelah diperluas ke wilayah bersalju, di mana visibilitas kamera dan perilaku jalan berbeda secara signifikan.

8.4 PERTIMBANGAN DESAIN SISTEM AI

Mendesain sistem AI bukan hanya tantangan teknis tetapi proses multidimensi yang mengintegrasikan penalaran etis, kendala peraturan, keterlibatan pemangku kepentingan, dan pandangan ke depan operasional. Proses desain yang cermat membantu memastikan bahwa sistem AI akan efektif, adil, aman, patuh, dan selaras dengan tujuan dan sasaran bisnis.

Bagian ini membahas aspek teknis dan proses bisnis dari desain. Tentu saja, setiap proses bisnis yang akan ditingkatkan oleh model AI perlu direvisi untuk mengatasi interaksi dengan model tersebut.

Mendefinisikan Tujuan dan Membingkai Masalah

Desain dimulai dengan kesengajaan dan kejelasan. Pemahaman mendalam tentang tujuan sistem AI memastikan bahwa pengembang dan pemangku kepentingan memiliki pemahaman yang sama tentang tujuan dan keterbatasannya. Singkatnya, ketika keberhasilan didefinisikan sebelumnya, perancang akan memiliki gagasan yang lebih jelas tentang bagaimana mencapainya. Beberapa pertanyaan yang perlu diajukan:

- Masalah bisnis atau operasional apa yang ingin dipecahkan oleh sistem?
- Apakah AI merupakan alat yang tepat untuk masalah ini?
- Apa hasil yang diinginkan dan tidak diinginkan dari sistem tersebut?

Sistem AI sering gagal karena masalah bisnis dibingkai dengan salah, atau asumsi yang tertanam tidak divalidasi. Fase desain adalah waktu yang tepat untuk mengungkap asumsi-

asumsi ini. Bahkan, fase desain adalah kesempatan terakhir untuk mempertanyakan semuanya sebelum sistem dibangun. Misalnya, sebuah lembaga pemerintah daerah ingin memprediksi di mana lubang jalan akan muncul. Membingkai ini sebagai masalah pemeliharaan prediktif, daripada tugas deteksi waktu nyata, mengarah pada desain model dan strategi data yang berbeda.

Pemilihan Arsitektur dan Model

Organisasi harus mempertimbangkan tidak hanya desain model AI tetapi juga arsitektur lengkap: bagaimana komponen berinteraksi, bagaimana aliran data, dan bagaimana hasil disampaikan. Pilihan arsitektur dan model harus didorong oleh tujuan, batasan, dan toleransi risiko.

Organisasi yang matang sering mengembangkan arsitektur logis dan arsitektur fisik, memungkinkan fungsi sistem didefinisikan pada tingkat konseptual, diikuti oleh desain lingkungan teknis.

Desain untuk Modularitas dan Keterlacakan

Sistem informasi modern sering kali modular, yang memfasilitasi peningkatan fleksibilitas dalam jangka panjang. Arsitektur modular mendukung skalabilitas, pemecahan masalah, pembaruan model, dan penjelasan. Praktik-praktik penting meliputi:

- Memisahkan pemasukan data, inferensi model, dan rendering output
- Mengaktifkan pengujian independen untuk setiap komponen
- Mencatat keputusan sistem internal dan transisi status
- Mendokumentasikan semua antarmuka tingkat modul dan sistem
- Melacak versi pipeline data, model, dan dependensi dalam registri untuk reproduksibilitas dan pengembalian
- Menetapkan kepemilikan dan tanggung jawab persetujuan perubahan untuk setiap modul untuk memastikan akuntabilitas

Memilih Tipe Model yang Tepat

Model AI harus sesuai untuk secara efektif mengatasi kompleksitas masalah bisnis, kebutuhan interpretasi, dan lingkungan data. Kecepatan evolusi model dan tipe AI membuat sulit untuk menghitung berbagai pilihan potensial, tetapi ini bukan wewenang tata kelola AI; sebaliknya, para ahli yang berkualifikasi harus mengeksplorasi jenis sistem AI kandidat dan membuat keputusan yang dapat dipertahankan dan diulang tentang jenis yang akan diimplementasikan.

Beberapa pertimbangan meliputi:

- Keputusan berbasis AI yang berisiko tinggi mungkin membutuhkan model yang lebih sederhana dan mudah dijelaskan.
- Lingkungan dengan keterbatasan sumber daya mungkin lebih menyukai model yang ringan atau berbasis cloud.
- Pilihan proprietary vs. open-source dapat memengaruhi transparansi.

Organisasi sering kali tertarik pada "hal-hal yang menarik perhatian," atau teknologi dan sistem baru yang ingin dimiliki oleh staf. Anggota staf sering kali dibutakan oleh bias yang berakar pada kebutuhan untuk tetap mengikuti perkembangan dan familiar dengan teknologi yang

muncul. Tanpa tindakan pencegahan yang tepat, organisasi mungkin akan memilih solusi berdasarkan "faktor kerennya" daripada kesesuaiannya yang sebenarnya.

Menghindari "Kecemburuan Model"

Model pembelajaran mendalam terbaru mungkin bukan solusi yang tepat untuk masalah bisnis spesifik suatu organisasi. Model yang lebih sederhana dapat mengungguli model yang kompleks ketika data terstruktur, fitur terdefinisi dengan baik, dan kemampuan menjelaskan sangat penting.

Setiap keputusan pemilihan model melibatkan pertimbangan timbal balik: tidak ada solusi sempurna. Keputusan harus didokumentasikan dan dibenarkan selama fase desain. Beberapa pertimbangan timbal balik model AI meliputi:

- ✓ Akurasi vs. interpretasi
- ✓ Kecepatan vs. biaya
- ✓ Keadilan vs. personalisasi
- ✓ Volume data vs. privasi
- ✓ Volume data vs. keamanan
- ✓ Minimisasi risiko vs. Inovasi

Misalnya, sebuah bank memilih model AI random forest untuk penilaian kredit karena memberikan keseimbangan yang baik antara akurasi dan kemampuan menjelaskan, yang sangat penting untuk kepatuhan terhadap peraturan.

Pengawasan Manusia dan Umpan Balik

Sistem AI dalam lingkungan sosio-teknis membutuhkan intervensi manusia untuk memainkan peran kunci dalam memantau, melakukan intervensi, dan meningkatkan hasil. Desain sistem dan proses yang baik mengantisipasi dan mendukung peran-peran ini.

Mendesain Proses dengan Manusia di Dalam Proses

Organisasi harus menentukan peran yang akan dimainkan manusia dan pada tahap mana dalam alur kerja proses mereka akan melakukan intervensi. Pilihan peran meliputi:

- ➢ **Penyetuju:** Pekerja harus menandatangani output AI sebelum tindakan dimulai.
- ➢ **Peninjau:** Pekerja meninjau output sistem audit secara berkala.
- ➢ **Titik eskalasi:** Sistem atau pihak yang terdampak menandai kasus-kasus yang tidak pasti untuk ditinjau oleh manusia.

Misalnya, sistem AI diagnostik perawatan kesehatan menandai kasus dengan skor kepercayaan rendah untuk ditinjau oleh ahli radiologi, sehingga meningkatkan keamanan, kepercayaan, dan keyakinan pada hasil.

Desain untuk Kegunaan dan Pemahaman

Pengawasan manusia hanya efektif jika mereka memahami apa yang dilakukan sistem dan peran mereka dalam operasi sistem yang sedang berlangsung. Pertimbangan meliputi:

- ❖ Penggunaan bahasa yang jelas, bukan skor kotak hitam
- ❖ Memberikan penjelasan yang jelas dan kontekstual
- ❖ Memungkinkan penggantian yang mudah dengan pencatatan alasan

Integrasi Umpan Balik

Umpan balik pengguna adalah sinyal berharga yang harus dimasukkan ke dalam peningkatan sistem. Organisasi harus mencari umpan balik dari pengguna sistem AI mereka dengan beberapa cara, termasuk:

- Survei daring
- Kelompok fokus
- Umpan balik eksplisit ("Keputusan ini salah")
- Sinyal implisit (misalnya, penggantian pengguna atau pengabaian alur kerja)
- Siklus tinjauan formal dengan pemangku kepentingan bisnis

Umpan balik dari pengguna harus memicu pelatihan ulang, perbaikan bug, atau tinjauan kebijakan dan tidak hanya dicatat dan diabaikan.

Metrik, Ambang Batas, dan Kontrol

Keputusan desain menentukan bagaimana keberhasilan dan kegagalan dapat diukur. Metrik harus mencerminkan tujuan bisnis, harapan pengguna dan/atau masyarakat, dan kendala operasional. Ambang batas kinerja harus dapat disesuaikan dan dapat dipertahankan, dan merupakan bagian dari desain, bukan ditentukan setelah penerapan.

Pemilihan Metrik

Masalah yang berbeda membutuhkan metrik yang berbeda. Metrik yang dipilih dengan buruk dapat menyebabkan perilaku yang tidak diinginkan atau organisasi tidak menyadari perkembangan penting.

Metrik adalah topik yang luas. Saya sarankan Anda membeli salinan buku *How to Measure Anything in Cybersecurity Risk*, Edisi ke-2; *Operational Risk: Measurement and Modelling;* atau *Key Performance Indicators: Developing, Implementing, and Using Winning KPIs,* Edisi ke-4, semuanya dari Wiley Publishing, untuk tinjauan komprehensif tentang metrik, untuk membantu pengembangan metrik.

Penetapan dan Penyesuaian Ambang Batas

Ambang batas dalam metrik menentukan kapan tindakan harus diambil. Ambang batas ini harus selaras dengan toleransi risiko, persyaratan peraturan, dan norma etika.

Praktik desain penting terkait ambang batas meliputi:

- Kalibrasi ambang batas berdasarkan dampak bisnis dan potensi kerugian
- Membutuhkan persetujuan formal untuk perubahan ambang batas
- Mendokumentasikan alasan untuk pengaturan yang dipilih

Kontrol Operasional

Organisasi perlu menerapkan kontrol khusus AI sebagai bagian dari penggunaan AI yang bertanggung jawab. Kontrol harus menjadi bagian dari desain sistem AI untuk pemantauan, failover, dan rollback sejak awal. Beberapa kontrol terkait AI mungkin termasuk:

- ✓ Peringatan untuk perilaku model yang tidak biasa
- ✓ Default failsafe (misalnya, tolak secara default)
- ✓ Penyebaran mode bayangan sebelum produksi
- ✓ "Tombol pemutus" untuk menonaktifkan model yang salah

Kontrol adalah topik besar lainnya: kontrol juga harus diimplementasikan di seluruh tumpukan operasional teknologi dan bisnis, untuk memastikan manajemen sistem informasi, personel, dan pihak ketiga yang baik. Buku-buku yang ditulis oleh penulis ini tentang sertifikasi CISM (*Certified Information Security Manager*) dan CRISC (*Certified in Risk and Information Systems Control*) berisi pembahasan, model, dan praktik yang komprehensif tentang pengendalian.

Kontrol dalam Undang-Undang AI Uni Eropa

Sistem AI berisiko tinggi harus mencakup mekanisme pengawasan manusia, pencatatan log, dan pemantauan pasca-implementasi yang semuanya bergantung pada keputusan yang dibuat selama fase desain.

Mengidentifikasi dan Mengelola Risiko Desain

Desain sistem AI memperkenalkan serangkaian risiko unik beberapa diwarisi dari pengembangan perangkat lunak tradisional, sementara yang lain berasal dari sifat probabilistik dan adaptif pembelajaran mesin. Risiko ini dapat bermanifestasi secara internal (misalnya, cacat desain, kualitas data yang buruk, atau kegagalan integrasi) atau eksternal (misalnya, penyalahgunaan, bias, atau kerugian sistemik). Mengidentifikasi dan mengelola risiko ini secara proaktif selama fase desain sangat penting untuk membangun sistem yang dapat dipercaya, tangguh, dan selaras dengan nilai-nilai organisasi.

Organisasi yang menunggu hingga sistem AI dibangun dan diterapkan berisiko mengalami gangguan melalui pengerjaan ulang yang mahal yang seharusnya dapat dihindari jika risiko telah diidentifikasi dalam proses desain atau sebelumnya.

Membingkai Risiko Desain AI

Tidak seperti sistem deterministik, sistem AI memperkenalkan nondeterminisme, ketidakjelasan, dan perilaku yang muncul, yang mempersulit identifikasi dan penilaian risiko. Sistem AI memperkenalkan risiko baru, termasuk:

- Perilaku yang tidak dapat diprediksi dalam kondisi data yang tidak terlihat
- Kesulitan melacak keputusan ke input atau logika tertentu
- Bias atau kerugian yang tertanam dalam data pelatihan, khususnya oleh variabel proksi
- Pergeseran perilaku model dari waktu ke waktu
- Ketergantungan berlebihan pada otomatisasi tanpa pengawasan yang tepat

Contoh sistem AI yang dilatih untuk memprediksi pengunduran diri karyawan mungkin secara tidak sengaja belajar untuk menghukum kandidat dari universitas yang lebih kecil atau latar belakang yang kurang terwakili karena bias data historis.

Membedakan Antara Sumber Risiko Internal dan Eksternal

Risiko internal (yang ada di dalam organisasi dan dalam kendalinya) muncul dari dalam desain sistem, arsitektur, proses pengembangan, lingkungan teknologi, dan proses bisnis terkait. Risiko-risiko ini dapat meliputi:

- Asumsi yang salah
- Dokumentasi yang tidak memadai
- Data pelatihan yang tidak memadai

- ❖ Integrasi yang buruk
- ❖ Pengujian yang tidak lengkap
- ❖ Kesenjangan kinerja model
- ❖ Kurangnya penjelasan atau transparansi
- ❖ Pemantauan yang tidak memadai untuk penyimpangan atau penyalahgunaan

Risiko eksternal (yang berada di luar organisasi dan sebagian besar di luar kendalinya) dapat meliputi:

- Penggunaan yang merugikan dan serangan siber
- Perubahan hukum atau peraturan
- Risiko dan pelanggaran rantai pasokan
- Reaksi budaya negatif atau ketidaksesuaian etika
- Penyalahgunaan pengguna atau sikap puas diri terhadap otomatisasi

Internal ≠ Terkendali

Meskipun organisasi menciptakan risiko internal, risiko tersebut tidak selalu sepenuhnya berada dalam kendalinya, terutama ketika menggunakan komponen pihak ketiga atau model yang telah dilatih sebelumnya.

Teknik Identifikasi Risiko Sistematis

Pendekatan standar dan terstruktur untuk identifikasi risiko mengurangi titik buta. Bagian ini membahas metode dan alat yang membantu tim secara sistematis mengungkap risiko desain.

Matriks Risiko

Salah satu alat yang paling banyak digunakan dalam manajemen risiko adalah matriks risiko, yang membantu memprioritaskan risiko berdasarkan probabilitas terjadinya dan potensi dampak atau kerugiannya. Teknik ini digunakan dalam keamanan siber dan manajemen risiko siber. Tabel 8.1 menggambarkan matriks risiko tipikal.

Tabel 8.1
Matriks Risiko Khas

Dampak Risiko	Probabilitas Rendah	Probabilitas Sedang	Kemungkinan Tinggi
Tingkat keparahan rendah	Memantau	Prioritas rendah	Prioritas sedang
Tingkat keparahan sedang	Prioritas rendah	Prioritas sedang	Prioritas tinggi
Tingkat keparahan yang tinggi	Prioritas sedang	Prioritas tinggi	Kritis

Misalnya, anggaplah model AI yang digunakan dalam pencitraan medis memiliki probabilitas rendah untuk salah mengklasifikasikan kondisi langka tetapi tingkat keparahannya tinggi jika terjadi kesalahan. Dalam hal ini, mungkin diperlukan pengawasan berlapis, bahkan jika kesalahan jarang terjadi.

Hierarki Mitigasi Risiko

Pendekatan berbasis hierarki memprioritaskan intervensi dari yang paling efektif hingga yang paling tidak efektif:

- **Hilangkan risiko** (misalnya, jangan gunakan AI dalam konteks berisiko tinggi)
- **Ganti dengan desain berisiko lebih rendah** (misalnya, gunakan model yang lebih sederhana dan mudah diinterpretasikan)
- **Rancang pengamanan** (misalnya, sertakan ambang batas kepercayaan, kontrol manusia dalam proses)
- **Latih dan pahami pengguna** (misalnya, sediakan dasbor interpretasi)
- **Terima dengan pemantauan** (misalnya, sertakan log audit, tinjauan pasca-pasar)

Ini adalah gambaran sederhana dari penanganan risiko, proses formal untuk menentukan apa yang harus dilakukan terhadap risiko yang teridentifikasi. Penanganan risiko merupakan komponen dari manajemen risiko, yang dibahas secara panjang lebar dalam buku-buku lain yang ditulis oleh penulis ini mengenai sertifikasi CISM (*Certified Information Security Manager*) dan CRISC (*Certified in Risk and Information Systems Control*).

Kompromi dalam Dunia Nyata

Eliminasi atau substitusi adalah konsep ideal tetapi seringkali tidak praktis karena tuntutan bisnis. Mitigasi risiko dalam AI seringkali melibatkan penataan kontrol secara strategis.

Pemetaan Pemangku Kepentingan dan Persepsi Risiko

Di seluruh organisasi, pemangku kepentingan yang berbeda memiliki persepsi risiko yang berbeda. Pemetaan peran, insentif, dan paparan pemangku kepentingan membantu mengidentifikasi vektor risiko yang terabaikan. Jenis risiko yang relevan bagi pemangku kepentingan meliputi:

- **Pengguna:** risiko kesalahan, kebingungan, ketergantungan berlebihan, halusinasi
- **Subjek:** risiko keputusan yang tidak adil, pelanggaran privasi
- **Operator:** risiko ketidaksesuaian beban kerja, tanggung jawab, biaya
- **Regulator:** risiko ketidakpatuhan, kurangnya transparansi
- **Publik:** masalah reputasi atau etika

Evaluasi Kasus Penggunaan dan Perencanaan Skenario

Kasus penggunaan dan lingkungan spesifik tempat sistem AI akan diterapkan sangat memengaruhi jenis dan tingkat keparahan risiko desain. Analisis skenario dan pengujian kasus ekstrem membantu mengungkap risiko spesifik konteks.

Klasifikasi Kasus Penggunaan dan Pemprofilan Risiko

Tidak semua kasus penggunaan diciptakan sama. Frekuensi dan ketelitian penilaian risiko harus proporsional dengan tingkat risiko. Beberapa faktor yang perlu dipertimbangkan meliputi:

- Dampak keputusan (rendah, sedang, tinggi)
- Paparan pengguna (internal vs. eksternal)

- Hak atau peluang yang terpengaruh
- Tingkat otomatisasi (penasihat vs. otonom)
- Regulasi (misalnya, Undang-Undang AI Uni Eropa, berbagai undang-undang privasi)
- Faktor kontekstual (norma budaya, pengawasan masyarakat, atau insiden sebelumnya di domain yang sama)

Misalnya, AI yang digunakan untuk rekomendasi Netflix mungkin memerlukan kontrol risiko yang lebih ringan daripada yang digunakan untuk keputusan kelayakan pengangguran.

Perencanaan Skenario dan Kejadian Buruk

Tim desain sistem AI dan pemangku kepentingan lainnya harus jujur bertanya: "Apa yang bisa salah?" Perencanaan skenario mencakup kegagalan yang diharapkan dan kasus-kasus ekstrem. Teknik-teknik yang digunakan meliputi:

- ✓ Pengujian tim merah dan pengujian lawan
- ✓ Latihan "Premortem" (membayangkan kegagalan dan menelusuri penyebabnya)
- ✓ Permainan peran penyalahgunaan dan serangan dengan beragam persona

Perancang alat kepolisian prediktif untuk contoh distrik sekolah membuat skenario di mana model tersebut menargetkan lingkungan tertentu secara berlebihan karena data historis yang bias, yang menyebabkan perubahan desain dan tinjauan eksternal.

Pembandingan dan Analisis Risiko Komparatif

Organisasi mungkin ingin membandingkan sistem AI yang mereka rencanakan untuk dikembangkan dengan tolok ukur yang ada atau implementasi serupa. Sumber potensial meliputi tolok ukur akademis seperti ImageNet atau GLUE, atau organisasi sejenis yang menggunakan sistem AI serupa. Pembandingan adalah alat lain yang membantu menginformasikan desain dan juga dapat membantu mempertahankan keputusan selama audit atau litigasi.

Pengujian, Percontohan, dan Validasi

Pengujian sistem AI bukanlah peristiwa tunggal: ini adalah alat desain. Pengujian sistem AI tidak dilakukan sekaligus tetapi secara bertahap sepanjang siklus pengembangan. Pengujian terstruktur dan uji coba terkontrol membantu memvalidasi asumsi, mengungkap risiko tersembunyi, dan menghasilkan basis bukti yang dibutuhkan untuk keputusan penerapan.

Pengujian Unit

Jika sistem AI merupakan bagian dari ekosistem, organisasi harus mempertimbangkan pengujian unit, di mana komponen individual diuji secara terpisah, untuk menentukan apakah komponen tersebut beroperasi seperti yang diharapkan dan sesuai desain.

Misalnya, sebuah koperasi kredit berencana untuk mengintegrasikan alat evaluasi pinjaman AI ke dalam sistem aplikasi pinjaman back-office-nya. Sebelum melakukan pengujian end-to-end, koperasi kredit memilih untuk menguji bagian-bagian dari sistem aplikasi pinjaman untuk memastikan sistem tersebut menghasilkan data yang benar untuk diteruskan ke sistem AI. Demikian pula, koperasi kredit juga mensimulasikan output sistem AI untuk menguji sistem aplikasi pinjaman untuk melihat apakah sistem tersebut merespons dengan benar.

Uji Coba Pra-Penerapan

Uji coba, yang merupakan peluncuran sistem produksi ke sejumlah pengguna terbatas untuk pengujian kesiapan akhir, mensimulasikan penerapan di dunia nyata dalam lingkungan yang terkontrol dan terbatas. Pertimbangan untuk pengujian percontohan meliputi:

- Mulai dengan pengguna yang "ramah"
- Tentukan kriteria keberhasilan dan kegagalan
- Sertakan beragam pengguna dan konteks
- Pantau perilaku yang tidak terduga
- Rencanakan tinjauan pasca-percontohan untuk menangkap masalah
- Tentukan masalah mana, jika ada, yang memerlukan perubahan pada sistem AI sebelum rilis umum

Misalnya, chatbot untuk dukungan pelanggan diuji dengan karyawan dan sukarelawan sebelum peluncuran penuh untuk mengidentifikasi kesenjangan dalam kualitas respons, logika eskalasi, dan nada.

Pengujian Mode Bayangan

Beberapa organisasi memilih untuk menerapkan sistem AI mereka dalam mode bayangan, mengintegrasikannya ke dalam proses bisnis yang ada untuk menangkap dan menganalisis output. Teknik ini mirip dengan pengujian paralel yang dilakukan dalam perencanaan pemulihan bencana TI, di mana sistem informasi cadangan diberi input yang sama dengan sistem produksi langsung untuk memverifikasi bahwa sistem cadangan berfungsi dengan benar.

Beberapa manfaat pengujian mode bayangan meliputi mendapatkan data kinerja langsung, memverifikasi bahwa semua integrasi berfungsi dengan benar, dan menentukan apakah sistem AI menghasilkan hasil yang sama dengan rekan manusianya.

Misalnya, sebuah bank yang menerapkan sistem peninjauan pinjaman AI menjalankannya dalam mode pengujian bayangan selama 60 hari, membandingkan saran AI dengan keputusan penjamin emisi yang sebenarnya. Perbedaan tersebut memberikan informasi untuk pelatihan ulang model dan peningkatan kemampuan penjelasan.

Pengujian Stres dan Pengujian Tim Merah

Pengujian stres dan pengujian tim merah sangat penting untuk memahami respons sistem AI terhadap input yang merugikan, kasus ekstrem, dan berbahaya. Jenis uji stres dapat mencakup:

- Injeksi data di luar distribusi
- Input yang merugikan yang dirancang untuk mengelabui model
- Simulasi pemadaman atau kehilangan data
- Simulasi pergeseran untuk menilai bagaimana model menangani perubahan distribusi input

Idealnya, organisasi telah melakukan uji stres dan red teaming, seperti pemodelan ancaman, sebelum pengembangan. Pemodelan ancaman adalah aktivitas umum dalam keamanan siber, di mana analis memeriksa persyaratan, arsitektur, dan/atau desain sistem untuk mengidentifikasi potensi kerentanan dan ancaman.

Integrasi Tata Kelola Risiko

Manajemen risiko fase desain harus terhubung dengan infrastruktur risiko dan kepatuhan organisasi yang lebih luas. Tanpa integrasi ini, bahkan risiko yang teridentifikasi dengan baik pun mungkin tidak diatasi. Organisasi akan lebih berhasil ketika berbagai disiplin manajemen risiko mereka diintegrasikan ke dalam satu proses yang luas.

Keselarasan dengan Kerangka Kerja Risiko Perusahaan

Organisasi harus melacak risiko desain AI dalam sistem dan format yang sama yang digunakan untuk risiko operasional lainnya. Ini termasuk penggunaan register risiko tunggal, rubrik penilaian umum, penetapan kepemilikan risiko, dan jalur eskalasi. Sebagai contoh, sebuah perusahaan telekomunikasi memasukkan risiko terkait AI ke dalam proses risiko keamanan informasi berbasis ISO/IEC 27005, memastikan perlakuan yang konsisten di seluruh proyek.

Mendokumentasikan Risiko Sisa dan Justifikasinya

Tidak semua risiko dapat dihilangkan. Ketika suatu organisasi mengidentifikasi risiko dan memutuskan untuk mengurangi risiko tersebut, seringkali ada risiko "sisa" (seperti kembalian dari uang Anda) yang dikenal sebagai risiko sisa. Risiko sisa harus didokumentasikan. Jika besarnya risiko sisa cukup besar, risiko tersebut harus diperlakukan sebagai risiko lain, seperti risiko lainnya, dan dianalisis, serta diputuskan apa yang harus dilakukan terhadapnya.

Jangan Sembunyikan Risiko Sisa

Risiko sisa tidak boleh disembunyikan. Regulator dan pemangku kepentingan seringkali lebih khawatir dengan risiko yang tidak diungkapkan daripada keberadaan risiko itu sendiri. Transparansi membangun kepercayaan.

Persiapan untuk Pemantauan Risiko Pasca-Penyebaran

Penilaian risiko tidak hanya tepat (bahkan bisa dibilang penting) dalam fase desain dan pengembangan sistem AI, tetapi juga sama pentingnya sepanjang masa pakai sistem AI. Oleh karena itu, pengendalian risiko desain harus diperluas hingga setelah penyebaran. Cara mempersiapkannya:

- Deteksi penyimpangan
- Ambang batas pelatihan ulang
- Dasbor pemantauan
- Alur kerja respons insiden, untuk berjaga-jaga
- Siklus umpan balik dari pengguna atau regulator

Misalnya, platform e-commerce suatu organisasi menggunakan deteksi anomali model AI untuk menandai perubahan kualitas rekomendasi atau keterlibatan pengguna setelah menerapkan algoritma peringkat baru.

8.5 PERTIMBANGAN MODEL PROPRIETARY

Organisasi yang memilih untuk mengembangkan dan menerapkan model AI proprietary menikmati kendali yang lebih besar tetapi juga harus menerima tanggung jawab

yang lebih besar. Tidak seperti penggunaan solusi AI pihak ketiga, model proprietary melibatkan kepemilikan ujung-ke-ujung atas persyaratan, desain, pelatihan, penerapan, dan semua hasil operasional dan hukum. Model AI proprietary, jika Anda dapat menemukan talenta untuk membangunnya, dapat menawarkan keunggulan kompetitif dan kemampuan kustomisasi, tetapi juga memperkenalkan kewajiban yang lebih tinggi, pengawasan peraturan, dan paparan tanggung jawab.

Kepemilikan dan Akuntabilitas

Organisasi yang membuat model AI proprietary dan menempatkannya ke dalam penggunaan produksi bertanggung jawab atas seluruh siklus hidup model, dari desain awal hingga pemeliharaan jangka panjang. Akuntabilitas meliputi:

- Keselarasan dan tata kelola etika
- Pengumpulan data dan legalitas
- Staf untuk merancang, membangun, dan mengoperasikan model
- Kinerja model, pengukuran keadilan, dan pemeliharaan
- Transparansi dan dokumentasi untuk digunakan dalam audit, proses hukum, dan tujuan lainnya
- Perilaku dan pemantauan pasca-implementasi
- Pembuangan sistem di akhir masa pakainya

Misalnya, penyedia layanan kesehatan yang membangun model diagnostiknya sendiri tidak dapat mengalihkan kesalahan kepada vendor jika model tersebut salah mendiagnosis suatu kondisi. Organisasi tersebut sepenuhnya bertanggung jawab atas hasil dan kemampuan audit. Namun, setelah analisis lebih lanjut, bahkan jika suatu organisasi menggunakan sistem AI pihak ketiga, organisasi tersebut tetap harus menerima sebagian besar akuntabilitas; bagaimanapun, organisasi tersebut telah menyeleksi dan memilih pihak ketiga tersebut. Singkatnya, akuntabilitas tidak dapat dialihdayakan.

Tanggung Jawab Hukum dan Paparan Perdata

Model kepemilikan dapat meningkatkan paparan terhadap tanggung jawab perdata dan peraturan, terutama jika penggunaan model tersebut menyebabkan kerugian. Tidak ada pihak lain yang dapat disalahkan.

Sumber risiko meliputi:

- Bias atau diskriminasi algoritmik
- Kerentanan keamanan
- Keputusan yang memengaruhi hak-hak yang dilindungi (misalnya, perumahan, pekerjaan, dan kredit)
- Kegagalan untuk memenuhi kewajiban hukum khusus industri
- Kekurangan dalam penjelasan atau dokumentasi

Undang-Undang AI Uni Eropa dan Model Proprietary

Berdasarkan Undang-Undang AI Uni Eropa, pengembang model proprietary berisiko tinggi menghadapi kewajiban yang lebih ketat, termasuk:

- Penilaian kesesuaian dan demonstrasi

- Dokumentasi teknis
- Pencatatan dan ketertelusuran
- Manajemen kualitas
- Mekanisme pengawasan manusia
- Pelaporan insiden

Dalam banyak kasus, hal-hal ini tidak dapat didelegasikan.

Klasifikasi Regulasi sebagai "Penyedia"

Dalam kerangka regulasi seperti Undang-Undang AI Uni Eropa, sebuah organisasi yang mengembangkan dan menerapkan model AI-nya sendiri biasanya dianggap sebagai "penyedia," bukan hanya "pengguna" (orang atau badan hukum yang menggunakan sistem AI di bawah wewenangnya sendiri). Hal ini menggeser beban kepatuhan ke hulu, yaitu kepada organisasi yang menjalankan model AI miliknya sendiri.

Tanggung jawab sebagai penyedia meliputi:

- ✓ Melakukan klasifikasi risiko
- ✓ Melakukan penilaian dampak
- ✓ Memelihara dokumentasi
- ✓ Memastikan keamanan siber dan ketahanan
- ✓ Memberikan pengungkapan dan pemberitahuan

Contoh perusahaan asuransi mengembangkan model penjaminan (underwriting) miliknya sendiri. Perusahaan tersebut harus memastikan kepatuhannya terhadap peraturan sektoral dan persyaratan hukum khusus AI, meskipun model tersebut tidak dijual atau digunakan oleh pihak lain.

Hak, Kualitas, dan Tata Kelola Data

Data merupakan bagian dari fondasi setiap model AI milik perusahaan. Ketika organisasi mengembangkan sistem AI mereka sendiri, mereka harus mengambil kendali penuh dan wewenang hukum atas data yang mereka gunakan, bersama dengan proses tata kelola yang baik untuk penanganannya.

Sumber Data dan Hak Hukum

Menggunakan data internal atau data yang diambil dari sumber lain tanpa dasar hukum yang tepat dapat membuat perusahaan menghadapi risiko yang signifikan, terutama ketika data tersebut tunduk pada peraturan privasi, seperti GDPR. Organisasi harus memastikan bahwa data dikumpulkan dengan persetujuan yang sesuai, dan hak penggunaan mencakup pelatihan AI; mereka juga harus mendokumentasikan dasar hukum untuk penggunaannya.

Data Internal ≠ Data Aman

Hanya karena data disimpan secara internal bukan berarti data tersebut dapat digunakan secara legal. Penggunaan kembali data tanpa pemberitahuan atau persetujuan yang tepat dapat melanggar prinsip minimalisasi data dan pembatasan tujuan.

Kualitas Data dan Integritas Pelabelan

Data berkualitas rendah atau bias dapat merusak kinerja dan meningkatkan risiko pengaduan dari warga serta potensi tindakan regulasi.

Kontrol terkait data untuk desain sistem AI meliputi:

- Menerapkan validasi data
- Mendokumentasikan metodologi pelabelan dan keandalan antar penilai
- Mencatat transformasi data dan langkah-langkah pra-pemrosesan

Misalnya, sebuah organisasi menerapkan model AI milik sendiri yang dirancang untuk mendeteksi cacat pada komponen industri. Model AI tersebut gagal dalam produksi karena pelabelan yang tidak konsisten di seluruh dataset pelatihan. Dokumentasi pedoman pelabelan dan audit tidak ada.

Integrasi Tata Kelola Data

Seperti yang disebutkan di bagian lain buku ini, tata kelola data merupakan prasyarat penting bagi organisasi yang bermaksud menerapkan sistem AI. Praktik tata kelola data memastikan penggunaan data yang aman, bertanggung jawab, dan sesuai. Praktik penting meliputi:

- Mendefinisikan pengelola data khusus AI
- Memberi tag pada dataset pelatihan dan validasi dalam katalog data
- Menerapkan kebijakan retensi dan penghapusan data
- Menyelaraskan dengan kerangka kerja etika data yang lebih luas

Privasi adalah komponen penting, yang terkait erat dengan tata kelola data. Kepatuhan terhadap hukum privasi yang berlaku (misalnya, GDPR, CPRA) memberlakukan persyaratan pada organisasi bahwa sistem AI mereka harus mendukungnya.

Penjelasan Model, Pengujian, dan Pemantauan

"Kami tidak tahu bagaimana atau mengapa sistem AI kami secara tidak adil menolak permohonan pinjaman untuk pelanggan kami," adalah pernyataan yang tidak ingin dibuat oleh organisasi mana pun secara publik. Tetapi jika model kepemilikan yang terlibat dalam pengambilan keputusan tidak dapat dijelaskan, cepat atau lambat organisasi akan mengalami masalah.

Model kepemilikan harus dapat dijelaskan kepada pengguna, dapat dipertahankan kepada regulator, dan dapat dipelihara dari waktu ke waktu. Organisasi yang membangun sistem mereka sendiri bertanggung jawab untuk menanamkan kualitas-kualitas ini sejak awal.

Kewajiban Penjelasan

Banyak kerangka peraturan (misalnya, GDPR, CPRA, Undang-Undang AI Uni Eropa) mensyaratkan bahwa keputusan otomatis yang memengaruhi individu harus dapat dijelaskan. Wajar untuk mengharapkan bahwa peraturan baru di yurisdiksi lain akan mengikuti hal yang sama. Pendekatan untuk penjelasan meliputi:

- Menggunakan model yang secara inheren dapat diinterpretasikan (misalnya, pohon keputusan, model linier)
- Menerapkan penjelasan post hoc (misalnya, SHAP, LIME)
- Menawarkan penalaran kontrafaktual atau berbasis kasus

- Memberikan justifikasi yang mudah dipahami pengguna dalam bahasa yang sederhana, ramah, tetapi dapat dipertahankan

Misalnya, sebuah perusahaan jasa keuangan menggunakan model AI penilaian kredit milik sendiri dan harus memberikan alasan penolakan kepada pelamar dengan cara yang akurat dan mudah dipahami.

Persyaratan Pengujian dan Validasi

Tidak seperti sistem pihak ketiga, yang diuji dan divalidasi oleh vendor, organisasi yang mengembangkan model milik sendiri harus mengandalkan tim internal mereka untuk merancang dan melaksanakan rezim pengujian. Rezim pengujian harus mencakup:

- Audit bias di seluruh atribut yang dilindungi
- Validasi kinerja pada skenario dunia nyata
- Pengujian ketahanan terhadap serangan (red teaming)
- Pengujian failover dengan campur tangan manusia

Pemantauan dan Pengendalian Perubahan

Model proprietary memerlukan manajemen insiden berkelanjutan dan proses formal untuk manajemen perubahan. Memang, seluruh praktik manajemen layanan TI (yang mencakup manajemen perubahan dan insiden, di antara beberapa praktik lainnya) sangat penting bagi setiap organisasi yang membangun dan mengoperasikan model AI proprietary. Praktik pemantauan meliputi:

- Menetapkan ambang batas peringatan untuk penurunan kinerja
- Mencatat input dan output untuk keterlacakan
- Mendeteksi pergeseran data
- Melatih ulang model secara teratur

Praktik pengendalian perubahan meliputi:

- Permintaan perubahan tertulis formal dengan alasan bisnis yang dinyatakan
- Persetujuan oleh pemangku kepentingan
- Rencana implementasi dan backout yang disetujui
- Implementasi alat manajemen konfigurasi untuk mencatat detail perubahan

Misalnya, pengecer melatih ulang mesin rekomendasi bertenaga AI-nya setiap bulan tetapi gagal mendokumentasikan perubahan hyperparameter. Akibatnya, perilaku pelanggan yang tidak terduga terjadi, dan regulator meminta bukti bahwa perubahan telah dinilai risikonya. Kesenjangan dalam dokumentasi tersebut mengungkap kelalaian perusahaan.

Sertifikasi ISO untuk Manajemen Layanan TI

Organisasi di industri yang diatur, serta mereka yang ingin menunjukkan kompetensi dalam operasi teknis mereka, mungkin ingin mengejar sertifikasi ISO/IEC 20000 untuk membuktikan manajemen layanan TI mereka. Perlu dicatat bahwa ISO/IEC 20000 tidak spesifik untuk AI tetapi berlaku secara umum untuk TI.

Hak Kekayaan Intelektual dan Risiko Kompetitif

Model AI milik perusahaan dapat memberikan keunggulan strategis, tetapi juga memperkenalkan risiko baru seputar hak kekayaan intelektual (HKI), kerahasiaan, dan potensi rekayasa balik atau penyalahgunaan.

Kepemilikan dan Perlindungan HKI

Organisasi, sebagai bagian dari tata kelola data secara keseluruhan, harus memelihara katalog yang akurat tentang kekayaan intelektual mereka. Praktik ini memungkinkan organisasi untuk menangani informasi tersebut dengan benar dengan risiko kompromi minimal. Organisasi harus menetapkan kepemilikan yang jelas atas aset AI dan data terkait.

Langkah-langkah yang perlu dipertimbangkan dalam perlindungan HKI meliputi:

- Apakah model AI dikembangkan secara internal, atau melalui kontraktor atau konsultan
- Apakah data pelatihan bersifat milik perusahaan, berlisensi, atau sumber terbuka
- Apakah bobot dan kode model didokumentasikan dan diamankan
- Apakah rahasia dagang dilindungi secara memadai

Paten vs. Rahasia Dagang

Model AI sering dilindungi sebagai rahasia dagang daripada dipatenkan, terutama ketika transparansi model dapat mengungkap kerentanan keamanan atau privasi. Pendekatan ini membatasi upaya hukum dalam beberapa kasus.

Rekayasa Balik dan Pencurian Model

Model AI milik perusahaan mungkin rentan terhadap rekayasa balik oleh pesaing atau pelaku jahat. Menurut penulis, perusahaan perangkat lunak komersial dan SaaS melakukan pekerjaan yang lebih baik dalam memastikan produk mereka tangguh daripada organisasi yang mengembangkan sistem secara internal. Intinya di sini bukanlah karena model AI tersebut milik perusahaan, tetapi karena sistem yang dikembangkan secara internal cenderung kurang tangguh daripada sistem yang diproduksi oleh vendor perangkat lunak.

Strategi mitigasi meliputi:

- ✓ Hanya gunakan API model, bukan kode sumber.
- ✓ Terapkan kontrol akses yang kuat pada semua API.
- ✓ Gunakan otentikasi, enkripsi, dan pengaburan untuk artefak model dan lalu lintas API jika memungkinkan.
- ✓ Pantau akses tidak sah, penyalahgunaan, dan serangan.
- ✓ Lakukan pengujian tim merah dan pemodelan ancaman secara berkala.
- ✓ Audit log penggunaan dan batasi laju kueri.
- ✓ Terapkan teknik yang menjaga privasi (privasi diferensial, pengacakan respons) dan penandaan output untuk mencegah pencurian model.

Sebuah contoh perusahaan fintech menemukan bahwa pesaingnya mampu menyimpulkan logika sistem deteksi penipuan berbasis AI miliknya melalui panggilan API berulang dan mereplikasi modelnya dengan sedikit modifikasi.

Penyalahgunaan Model dan Kontrol Akses

Organisasi perlu mengontrol siapa yang dapat menggunakan atau memodifikasi model milik mereka. Kontrol akses yang kuat (yang merupakan praktik luas dan penting) sangat penting untuk mencegah penyalahgunaan atau kerusakan yang tidak disengaja. Pertimbangan meliputi:

- Menerapkan konsep "perlu tahu" dan "hak akses minimal".
- Menerapkan izin berbasis peran untuk pengembangan dan penerapan.
- Mencatat perubahan pada bobot atau logika model.
- Memanfaatkan alur kerja persetujuan internal untuk pembaruan model.
- Menerapkan prosedur pencabutan untuk model yang tidak digunakan atau yang telah dikompromikan.
- Secara berkala meninjau akses istimewa ke sistem AI.

Sumber Daya dan Kesiapan Organisasi

Seperti upaya pengembangan sistem internal lainnya, pengembangan model milik sendiri merupakan proses yang membutuhkan banyak sumber daya dan investasi berkelanjutan. Organisasi harus memastikan mereka memiliki infrastruktur, talenta, dan tata kelola untuk mendukung AI milik sendiri dalam skala besar.

Banyak organisasi meremehkan tingkat upaya yang dibutuhkan untuk membangun dan mempertahankan model AI milik mereka sendiri. AI tidak unik dalam hal ini, karena seni memperkirakan proyek perangkat lunak telah menjadi tantangan bagi organisasi sejak tahun 1960-an. Kita tidak lebih baik dalam memperkirakan tingkat upaya untuk proyek-proyek besar dan kompleks yang menggunakan teknologi baru daripada sebelumnya.

Infrastruktur Teknis

Model AI kustom biasanya membutuhkan infrastruktur yang lebih menuntut daripada solusi siap pakai. Organisasi harus berhati-hati dan mengelola pengeluaran proyek dengan cermat. Dan dibutuhkan keberanian khusus untuk menghentikan proyek AI jika sudah melenceng dari jalur yang seharusnya.

Kebutuhan infrastruktur meliputi:

- Kluster komputasi GPU/TPU
- Pipeline data yang dapat diskalakan
- Lingkungan hosting model yang aman
- Sandbox untuk pengembangan dan pengujian
- Lingkungan pengembangan dan pengujian yang terpisah

Pengembangan Bakat dan Keterampilan

Untuk berhasil melaksanakan proyek pengembangan sistem skala besar, organisasi membutuhkan anggota tim tingkat senior untuk memimpin upaya tersebut. Pengembangan internal meningkatkan ketergantungan pada bakat khusus, termasuk ilmuwan data, insinyur ML, dan profesional tata kelola model. Karena tingginya permintaan akan talenta AI, keterampilan ini langka dan harganya pun sesuai. Melatih anggota tim internal yang tepat dapat membantu mengisi kesenjangan ini.

Ide manajemen talenta meliputi:

- Strategi retensi dan kompensasi untuk staf teknis kunci
- Pelatihan silang antara tim AI dan kepatuhan
- Transfer pengetahuan internal dan dokumentasi
- Pembentukan "pusat keunggulan" AI untuk mendorong kolaborasi dan pembelajaran

Risiko Titik Kegagalan Tunggal
Ketika hanya satu atau dua individu yang memahami cara kerja model kepemilikan (dan sistem yang dikembangkan secara internal lainnya), organisasi akan menghadapi tantangan keberlanjutan dan manajemen risiko jika individu-individu tersebut keluar.

Kapasitas Tata Kelola dan Tinjauan Etika

Organisasi yang mengembangkan model AI milik sendiri perlu membangun alur kerja tata kelola tambahan. Mengingat pengembangan perangkat lunak tradisional dan pengembangan model AI memiliki kesamaan, SDLC (Software Development Lifecycle) yang matang mungkin dapat diadaptasi untuk pengembangan model AI, daripada membangun proses siklus hidup pengembangan yang serupa secara terpisah dari awal. Perhatikan batasan-batasan berikut:

- Terapkan dewan persetujuan untuk model internal.
- Gunakan kriteria tinjauan etika yang diperluas.
- Integrasikan tata kelola AI dengan fungsi hukum, risiko, dan audit.
- Tetapkan "pemilik model" sebagai pihak yang bertanggung jawab atas pengawasan berkelanjutan.

Misalnya, perusahaan rintisan teknologi menambahkan Dewan Tata Kelola AI formal untuk meninjau model milik sendiri melalui setiap tahap pengembangan, termasuk persetujuan dari pimpinan bidang hukum, keamanan, dan etika.

8.6 RINGKASAN

Merancang sistem AI yang andal dan tangguh secara bertanggung jawab membutuhkan dokumentasi komprehensif dan pemahaman kontekstual di seluruh siklus hidup. Dari konsep awal hingga pengawasan pasca-implementasi, setiap fase memperkenalkan keputusan yang membentuk kinerja, postur kepatuhan, dan dampak sosial. Pendekatan terstruktur untuk mendesain sistem AI, yang menekankan dokumentasi siklus hidup, keselarasan bisnis, kesadaran risiko, dan kontrol yang disesuaikan untuk model kepemilikan, membantu memastikan hasil yang menguntungkan.

Mendokumentasikan siklus hidup AI sangat penting untuk ketertelusuran, kepatuhan, dan audit di masa mendatang. Tahap desain dan pembangunan harus mencakup pernyataan yang jelas tentang tujuan sistem AI, peran pemangku kepentingan, sumber data, arsitektur model, dan rasionalisasi keputusan. Sebelum pelatihan dan pengujian, dokumentasi harus mencakup asal usul dataset, langkah-langkah pra-pemrosesan, metrik kinerja, dan upaya mitigasi risiko. Ketika data pelatihan mencakup data privasi, organisasi harus secara formal menetapkan dasar hukum untuk pemrosesan. Catatan yang menyeluruh tidak hanya

membantu memvalidasi hasil sistem tetapi juga memberikan dasar untuk penjelasan dan pembelaan peraturan sepanjang masa operasional model. Hasil dari sistem AI, sejauh mungkin, harus ditentukan sebelum pengembangan.

Desain yang efektif juga dimulai dengan pemahaman yang jelas tentang konteks bisnis dan kasus penggunaan yang dimaksudkan. Organisasi harus mengidentifikasi masalah apa yang dipecahkan oleh sistem AI, siapa yang dilayaninya, dan bagaimana outputnya terintegrasi ke dalam alur kerja operasional. Menyelaraskan kemampuan sistem dengan tujuan bisnis, ambang batas kinerja, ketersediaan data, dan pertimbangan etis sangat penting untuk keberhasilan. Keterlibatan pemangku kepentingan sangat penting, karena pengguna, populasi yang terdampak, dan badan pengawas masing-masing memberikan wawasan berharga tentang kendala, risiko, dan konsekuensi yang tidak diinginkan. Kasus penggunaan yang diartikulasikan dengan baik memastikan bahwa pilihan teknis selaras dengan kendala dunia nyata dan memberikan hasil yang terukur.

Penilaian dampak berfungsi sebagai tonggak penting dalam proses desain, membantu organisasi mengungkap potensi bahaya, kesenjangan kepatuhan, atau titik buta etis sebelum sistem AI diimplementasikan. Penilaian ini mengevaluasi faktor-faktor seperti bias, kemampuan menjelaskan, otonomi, dan keadilan di seluruh kasus penggunaan. Proses penilaian dampak yang terstruktur harus melibatkan beragam pemangku kepentingan, mengevaluasi strategi mitigasi, dan mendokumentasikan risiko yang diterima, diselesaikan, dan residual.

Temuan dari penilaian ini harus menjadi dasar revisi desain, pilihan antarmuka pengguna, pengamanan keterlibatan manusia, dan kebijakan penerapan. Di yurisdiksi seperti Uni Eropa dan Kanada, penilaian ini semakin menjadi persyaratan hukum. Semua kegiatan ini harus didokumentasikan.

Mendesain untuk keadilan, keamanan, kepercayaan, dan ketahanan membutuhkan lebih dari sekadar niat baik; hal ini melibatkan kebijakan dan praktik terbaik khusus di seluruh fase desain. Praktik-praktik ini mencakup pengumpulan persyaratan yang menyeluruh, pemilihan arsitektur model, definisi metrik, desain pengawasan manusia, dan integrasi umpan balik. Sistem harus dievaluasi tidak hanya untuk kinerja teknis tetapi juga untuk keselarasan dengan nilai-nilai organisasi dan keberlanjutan operasional. Desain harus mengantisipasi kebutuhan akan pemantauan, pelatihan ulang, dan penyesuaian yang berkelanjutan, menanamkan kontrol tersebut ke dalam sistem sejak awal. Saat desain bergerak melalui fase-fasenya, persetujuan formal harus digunakan untuk menjaga proyek tetap sesuai rencana.

Organisasi yang mengembangkan model kepemilikan menghadapi serangkaian kewajiban dan risiko yang lebih tinggi. Kepemilikan ujung-ke-ujung berarti mereka sepenuhnya bertanggung jawab untuk memastikan kepatuhan hukum, mencapai hasil etis, dan memastikan penjelasan model. Pengembangan perangkat lunak berpemilik juga menimbulkan kekhawatiran seputar hak data, kekayaan intelektual, infrastruktur, dan kemampuan tenaga kerja. Tata kelola AI harus lebih kuat, pengujian lebih menyeluruh, dan dokumentasi lebih lengkap dan tepat. Dalam banyak rezim peraturan, termasuk Undang-Undang AI Uni Eropa, pengembang model berpemilik dianggap sebagai "penyedia," dan harus memenuhi

persyaratan pelaporan, keterlacakan, dan pengawasan yang lebih ketat. Organisasi dengan SDLC (*Software Development Life Cycle*) yang matang mungkin dapat memperluasnya untuk mencakup pengembangan model AI berpemilik.

BAB 9
TATA KELOLA DATA DAN PELATIHAN MODEL

9.1 TATA KELOLA DATA

Sistem AI hanya akan andal dan tangguh sejauh data yang digunakan untuk melatihnya. Karena alasan ini, tata kelola data merupakan pendorong penting (bahkan prasyarat) pengembangan AI yang bertanggung jawab. Meskipun sistem AI secara teknis dapat berfungsi tanpa struktur tata kelola formal, kepercayaan, akuntabilitas, integritas, dan efektivitasnya akan terganggu. Dan, banyak organisasi yang sama sekali tidak mempraktikkan tata kelola data atau menerapkannya dengan ceroboh. Kekurangan ini menciptakan kerentanan, termasuk kumpulan data pelatihan yang bias atau tidak akurat, ketidakpatuhan terhadap kewajiban hukum, dan kegagalan integritas data, di antara lainnya. Kerangka kerja tata kelola data yang kuat dan efektif membantu mengurangi risiko ini dan menetapkan fondasi yang lebih kokoh untuk AI yang etis, legal, dan berkinerja tinggi.

Komponen utama tata kelola data meliputi:

- Klasifikasi data: Skema penetapan tingkat risiko pada kumpulan data, bersama dengan persyaratan penanganan dan tata kelola untuk setiap tingkat
- Diagram aliran data (DFD): Dikembangkan oleh arsitek data atau sistem, ini adalah representasi visual dari aliran data di dalam dan di antara sistem dan entitas
- Kontrol input: Untuk memastikan bahwa data yang mengalir ke dalam sistem disetujui oleh manajemen dan juga diformat dengan benar
- Kontrol output: Untuk memastikan bahwa data yang mengalir keluar dari sistem disetujui oleh manajemen dan diformat dengan benar
- Pencegahan kehilangan data (DLP): Perangkat lunak yang memantau pergerakan data, untuk memberi peringatan atau memblokir pergerakan tertentu jika melanggar kebijakan

Bagian ini memberikan tinjauan singkat tentang tata kelola data dan penerapannya pada sistem AI. Untuk tinjauan komprehensif tentang tata kelola data perusahaan, lihat Panduan Praktisi untuk Mengoperasionalkan Tata Kelola Data atau Tata Kelola Data untuk Pemula dari Wiley Publishing.

Mendefinisikan Tata Kelola Data untuk AI

Tata kelola data mengacu pada kerangka kerja pengelolaan data perusahaan melalui kebijakan, proses, alat, peran, dan tanggung jawab untuk memastikan data akurat, dapat digunakan, aman, dan sesuai dengan persyaratan yang relevan dan peraturan yang berlaku. Dalam pengembangan AI, tata kelola memastikan bahwa dataset pelatihan, pengujian, dan operasional tidak hanya sesuai secara teknis tetapi juga sah secara hukum dan etika serta akan digunakan dengan tepat.

Tata kelola data harus mencakup isu-isu seperti siapa pemilik data, bagaimana data tersebut diperoleh, apakah data tersebut dapat digunakan untuk pengembangan AI,

bagaimana data tersebut harus dilindungi, dan apa yang diatur oleh hukum, peraturan, dan persyaratan yang berlaku untuk perlindungan dan penggunaan yang tepat. Tanpa pengamanan ini, organisasi berisiko membangun sistem AI di atas fondasi yang salah atau melanggar hukum.

Otoritas Hukum dan Penggunaan Data yang Sah

Langkah pertama dalam tata kelola data dalam konteks AI adalah konfirmasi bahwa organisasi memiliki hak hukum yang dapat dipertahankan untuk menggunakannya.

Mengkonfirmasi Hak Hukum untuk Menggunakan Data

Organisasi harus secara formal memverifikasi bahwa mereka berwenang untuk mengumpulkan, memproses, menyimpan, dan menganalisis data yang akan dimasukkan ke dalam sistem AI. Hak-hak ini dapat didasarkan pada satu atau lebih hal berikut:

- Persetujuan langsung pengguna (misalnya, persetujuan melalui kotak centang di aplikasi)
- Persetujuan implisit pengguna (termasuk dalam ketentuan penggunaan; persetujuan implisit tidak cukup di beberapa yurisdiksi, termasuk Uni Eropa melalui GDPR)
- Ketentuan kontraktual (misalnya, perjanjian berbagi data)
- Kebijakan privasi atau pemberitahuan deklarasi praktik privasi
- Ketentuan peraturan (misalnya, berbagi data kesehatan yang sesuai dengan HIPAA, yang seringkali memerlukan deidentifikasi dan perjanjian rekan bisnis (BAA))

Tinjauan formal oleh penasihat hukum atas semua hal ini, bersama dengan pernyataan tertulis, dianggap penting bagi suatu organisasi untuk melanjutkan penggunaan data untuk melatih sistem AI.

Persyaratan Dokumentasi

Setelah akses yang sah ditetapkan, organisasi harus mendokumentasikannya sedetail mungkin sesuai dengan peraturan, kebijakan, atau budaya organisasi. Kebijakan tata kelola data harus mencakup:

- ✓ Pendaftaran sumber data
- ✓ Catatan persetujuan (jika berlaku)
- ✓ Referensi kontrak
- ✓ Referensi kebijakan privasi/pemberitahuan praktik privasi
- ✓ Pembenaran hukum khusus yurisdiksi

Misalnya, sebuah perusahaan teknologi medis membeli kumpulan data yang berisi catatan pasien yang telah dianonimkan untuk pemodelan prediktif. Beberapa bulan kemudian, investigasi pemerintah mengungkapkan bahwa data tersebut dapat diidentifikasi kembali dalam kondisi tertentu.

Perusahaan tersebut belum memverifikasi proses anonimisasi vendor atau dasar hukum untuk menjual data tersebut. Proyek tersebut dihentikan, dan denda yang besar dikenakan.

Risiko Penggunaan Sekunder

Beberapa organisasi menggunakan kembali data di luar tujuan pengumpulan dan persetujuan aslinya. Organisasi yang kurang memiliki tata kelola data kemungkinan akan

mengalami penyimpangan fungsi, yang menyebabkan pelanggaran kebijakan privasi, kontrak, atau peraturan. Dan tanpa tata kelola data, semua ini dapat terjadi dengan sedikit atau tanpa kesadaran. Proses peninjauan dan penilaian dampak yang dilakukan sebagai bagian dari tata kelola data harus mengidentifikasi penggunaan sekunder berisiko tinggi.

Saat Menggunakan Sistem AI Pihak Ketiga

Organisasi yang berencana menggunakan sistem AI pihak ketiga harus mempertimbangkan beberapa faktor tambahan. Secara khusus, ketika suatu organisasi melatih sistem AI pihak ketiga dengan informasi pribadi, organisasi tersebut harus secara eksplisit memahami apakah data pelatihan tersebut akan digunakan di tempat lain. Tergantung pada tingkat risikonya, organisasi dalam posisi ini mungkin ingin memberlakukan ketentuan khusus dalam perjanjian hukum, termasuk:

- Data pelatihan tidak boleh digunakan untuk sistem, pelanggan, atau tujuan lain.
- Organisasi berhak untuk mengaudit organisasi pihak ketiga dan sistem terkait.
- Audit dan penilaian keamanan siber harus dilakukan oleh pihak ketiga yang berkualifikasi.
- Organisasi berhak untuk mengakhiri perjanjian tanpa alasan.
- Vendor harus setuju untuk menghancurkan data atas permintaan dan pada saat pengakhiran perjanjian.

Topik ini dibahas lebih detail di Bab 11.

Menilai Kualitas dan Integritas Data

Meskipun menetapkan dasar hukum untuk data sangat penting, data yang diperoleh secara sah masih dapat merusak sistem AI jika data tersebut kurang berkualitas atau berintegritas.

Dimensi Kualitas Data

Kebijakan tata kelola data harus mendefinisikan kualitas data dan menguraikan proses untuk mengevaluasinya. Dimensi utama meliputi:

- **Akurasi:** Apakah nilai-nilai yang terdapat dalam dataset benar dan bebas dari kesalahan?
- **Kelengkapan:** Apakah semua kolom yang diperlukan telah terisi?
- **Integritas:** Apakah hubungan antar dataset dipelihara dengan benar?
- **Ketepatan waktu:** Apakah data mutakhir?
- **Konsistensi:** Apakah nilai-nilai mengikuti format dan aturan yang diharapkan?

Kesesuaian dengan Tujuan Berarti Lebih dari Sekadar Akurasi

Model AI dapat secara teknis sangat baik tetapi tetap gagal jika tidak selaras dengan tujuan penggunaannya. Kesesuaian dengan tujuan mencakup kepatuhan terhadap peraturan, integrasi dengan alur kerja, dan kesesuaian untuk lingkungan operasional. Sistem AI harus memberikan nilai dalam konteks penggunaannya, bukan hanya dalam hasil pengujian.

Mendeteksi dan Mencegah Pergeseran Data

Untuk model yang diperbarui atau dipelajari dari waktu ke waktu, tata kelola harus mencakup mekanisme untuk mendeteksi dan menanggapi pergeseran data, yang terjadi ketika data keluaran baru menyimpang secara signifikan dari distribusi aslinya. Protokol pemantauan harus diwajibkan sebagai bagian dari tata kelola siklus hidup model AI.

Memastikan Data yang Cukup dan Representatif

Kecukupan data mengacu pada memiliki volume data yang cukup untuk melatih model secara efektif. Representativitas mengacu pada memastikan bahwa data secara akurat menangkap keragaman yang relevan dari lingkungan tempat AI akan beroperasi.

Menghindari Underfitting dan Overfitting

Volume data pelatihan yang tidak mencukupi dapat menyebabkan underfitting, di mana model tidak mampu melakukan generalisasi secara efektif. Terlalu banyak data yang sempit atau berulang dapat menyebabkan overfitting, di mana model mempelajari pola yang tidak relevan. Tata kelola data yang efektif membantu menemukan keseimbangan yang tepat dengan mendefinisikan ambang batas kecukupan dan batasan penggunaan kembali.

Merepresentasikan Populasi Dunia Nyata

Untuk menghindari munculnya bias, kerangka kerja tata kelola AI harus memasukkan pemeriksaan keadilan untuk mendeteksi dan mengatasi kesenjangan representasi. Contohnya termasuk audit demografis, pemeriksaan distribusi data (seperti distribusi geografis), dan deteksi outlier.

Sebagai contoh, sebuah organisasi menerapkan alat penyaringan resume berbasis AI dan melatihnya dengan data perekrutan historis; sistem AI tersebut sangat mengutamakan kandidat laki-laki untuk peran teknis. Audit pasca-implementasi mengungkapkan bahwa data pelatihan kurang mewakili pelamar perempuan. Tata kelola dapat mendeteksi bias tersebut lebih awal jika mereka mencarinya.

Evaluasi Kesesuaian Tujuan

Agar sistem AI berhasil, data yang digunakan untuk pelatihannya tidak hanya harus bersih dan representatif, tetapi juga harus sesuai dengan tugas dan konteks model AI. Tinjauan tata kelola harus mengevaluasi apakah dataset mendukung tujuan model AI. Misalnya, menggunakan data sentimen media sosial untuk memprediksi perilaku pemungutan suara mungkin tidak tepat tanpa validasi silang terhadap jajak pendapat atau kontrol demografis. Mengambil data yang ditujukan untuk satu tujuan dan menggunakannya untuk tujuan yang berbeda harus didekati dengan hati-hati dan harus melalui tinjauan dan persetujuan.

Pelatihan AI ≠ Penggunaan Umum

Kumpulan data yang digunakan untuk melatih sistem AI dalam penerjemahan bahasa mungkin tidak cocok untuk melatih chatbot dalam konteks pemberian nasihat medis. Bahkan data yang secara teknis serupa pun dapat berbeda dalam implikasi etis dan risiko hukum.

Tata Kelola Data dan Peran

Tata kelola data yang efektif membutuhkan peran dan tanggung jawab yang jelas. Dua peran utama adalah:

- **Pengelola data:** Pengelola ini adalah penjaga operasional yang bertanggung jawab untuk memastikan kualitas data, dokumentasi, dan penyelesaian masalah. Mereka memastikan bahwa kebijakan tata kelola diterapkan secara konsisten di seluruh siklus hidup data.
- **Pemilik data:** Pemilik bertanggung jawab atas penggunaan strategis data. Mereka menyetujui kasus penggunaan, mengesahkan penggunaan yang sah, dan menandatangani penilaian risiko. Kerangka kerja tata kelola harus mengidentifikasi pemilik data berdasarkan domain atau kumpulan data.

Semua aktivitas ini harus menjadi bagian dari proses bisnis yang terdokumentasi dan tercatat. Pemilik dan pengelola data harus diidentifikasi dengan bijak, dan masing-masing harus memiliki pemahaman yang jelas tentang tanggung jawab mereka.

Dalam contoh perusahaan asuransi, departemen penjaminan memiliki data aktuaria, sementara departemen pemasaran memiliki kumpulan data segmentasi pelanggan. Tanpa kepemilikan yang jelas, tinjauan tata kelola akan terhenti atau tidak efektif, dan persetujuan risiko akan menjadi hambatan.

Keamanan, Privasi, dan Penggunaan Etis

Agar efektif, program tata kelola data harus terintegrasi dengan keamanan siber dan privasi, sehingga keduanya beroperasi sebagai struktur yang kohesif untuk melindungi data dan memastikan penggunaannya yang bertanggung jawab dan legal. Pengenalan AI memperkuat kebutuhan ini, dengan penggunaan data yang baru, di mana pelanggaran dan penyalahgunaan dapat berkembang dengan cepat.

Penilaian Risiko Privasi

Program tata kelola data harus mewajibkan penilaian dampak privasi (PIA) untuk semua sistem AI yang akan memproses data pribadi. Penilaian ini mengevaluasi risiko kepatuhan, identifikasi ulang, inferensi, dan hasil diskriminatif.

Kontrol Keamanan

Berbagai macam kontrol harus diimplementasikan untuk memastikan perlindungan yang memadai dan penggunaan yang tepat dari kumpulan data terkait AI. Pada lapisan data, kontrol seperti enkripsi, manajemen akses, dan pencatatan audit harus diwajibkan untuk setiap kumpulan data terkait AI, termasuk data pelatihan. Pada lapisan infrastruktur, serangkaian kontrol yang komprehensif harus diwajibkan untuk melindungi fasilitas, jaringan, sistem, dan sistem manajemen basis data dari serangan dan penyalahgunaan. Buku-buku karya penulis ini tentang sertifikasi keamanan siber, termasuk CISSP, CISM, CISA, dan CRISC, membahas topik kontrol secara mendalam.

Pertimbangan Etis

Karena penggunaan data yang baru dan berkembang dalam sistem AI, organisasi perlu memasukkan pertimbangan etis ke dalam program tata kelola data mereka. Tinjauan etis harus membahas isu-isu di luar yang terkait dengan kebijakan, persyaratan, dan peraturan. Tinjauan

etis dapat dianggap sebagai langkah persetujuan akhir, karena penting untuk mempertimbangkan bahwa meskipun penggunaan data tertentu mungkin diizinkan dan legal, hal itu mungkin tetap tidak bijaksana. Contoh hal-hal yang perlu diperhatikan dalam tinjauan etis meliputi:

- Data pelatihan yang menyematkan bias rasial, gender, atau sosial ekonomi melalui variabel proksi
- Data pelatihan yang dikumpulkan dalam kondisi eksploitatif atau tidak jujur
- Data pelatihan yang melanggar privasi, meskipun secara teknis dianonimkan atau dipseudonimkan

Sebagai contoh, pertimbangkan masalah hukum kontroversial yang melibatkan model AI yang dilatih pada forum terapi online yang di-scrape. Data tersebut bersifat publik, tetapi penggunaannya melanggar harapan pengguna dan memicu reaksi negatif. Pesan moral dari cerita ini: penggunaan kumpulan data tertentu dapat sah secara hukum namun tidak etis.

Kedewasaan dan Implementasi Organisasi

Tata kelola data bukanlah tugas kepatuhan sekali saja, melainkan proses siklus hidup yang berlanjut sepanjang masa pakai sistem AI—bahkan, sepanjang masa pakai penggunaan data oleh organisasi. Agar efektif dan memenuhi kebutuhan semua pihak, tata kelola data membutuhkan dukungan kepemimpinan dan dukungan institusional, dan bahkan dapat memicu perubahan budaya.

Organisasi dapat memilih untuk memusatkan tata kelola data di bawah satu badan atau menggabungkannya di berbagai departemen. Meskipun modelnya akan bervariasi, semuanya membutuhkan dukungan eksekutif, pelatihan dan kesadaran, metrik, prosedur yang terdokumentasi, dan catatan bisnis resmi.

Bagi organisasi yang membangun program tata kelola data baru di mana sebelumnya tidak ada, disarankan agar mereka meniru struktur tata kelola yang ada, meminjam apa yang berhasil, dan mempertimbangkan penyesuaian yang diperlukan untuk memastikan keberhasilan. Bagi organisasi yang menggunakan platform GRC atau sistem serupa untuk mengelola proses tata kelola, disarankan agar tata kelola data juga menggunakan sistem ini untuk konsistensi dan kemudahan penggunaan.

Organisasi yang berfokus pada kematangan proses harus mempertimbangkan untuk mengadopsi model kematangan, untuk membantu mereka menilai dan meningkatkan postur tata kelola mereka. Tingkatannya biasanya meliputi:

- **Ad hoc:** Dokumentasi atau pengawasan minimal
- **Berkembang:** Beberapa peran formal dan kebijakan dasar
- **Terdefinisi:** Tanggung jawab yang jelas dan prosedur standar
- **Terkelola:** Pemantauan, pengukuran, dan pengambilan keputusan berbasis risiko
- **Dioptimalkan:** Pemantauan berkelanjutan, otomatisasi, metrik, dan pengambilan keputusan yang disesuaikan dengan risiko

Sebuah perusahaan pengiriman barang dengan tata kelola yang ad hoc meluncurkan sistem penentuan rute pengiriman berbasis AI. Ketika kegagalan pengiriman meningkat, dibutuhkan waktu berminggu-minggu bagi organisasi untuk melacak masalah tersebut hingga ke data

lokasi yang rusak. Program tata kelola yang matang dapat mendeteksi masalah tersebut lebih cepat, bahkan hanya dalam hitungan jam.

Jebakan dan Mode Kegagalan Umum

Bahkan organisasi yang terorganisir dengan baik dan matang pun dapat kesulitan dalam implementasi. Organisasi yang kurang matang tidak akan mampu menghindari serangkaian masalah. Berikut beberapa contohnya:

- **Memperlakukan tata kelola sebagai latihan kepatuhan:** Tata kelola yang hanya berfokus pada pemenuhan persyaratan, daripada memungkinkan penggunaan data yang dapat dipercaya, seringkali menyebabkan resistensi dan kontrol yang tidak efektif.
- **Terlalu terpusat:** Struktur tata kelola yang terlalu terpusat, terutama di organisasi yang lebih besar, dapat menciptakan hambatan dan menyebabkan disintermediasi tata kelola data. Keseimbangan yang tepat harus dicapai antara kontrol dan kelincahan.
- **Kelebihan alat:** Penggunaan alat dan dasbor tata kelola yang berlebihan dapat membebani tim, menyebabkan penggunaan yang tidak konsisten. Kesederhanaan dan pelatihan adalah kunci untuk praktik tata kelola data yang dapat diterapkan.
- **Menggunakan salinan bayangan data pelatihan:** Suatu organisasi mengekstrak data produksi dari aplikasi bisnis dan melakukan penyaringan ad hoc secara offline tanpa mencatat prosedurnya. Tata kelola data (jika ada) akan buta terhadap penggunaan data ini.

Manfaat Tata Kelola dalam Pengembangan AI

Jika dirancang dan diimplementasikan dengan baik, tata kelola data tidak hanya mengurangi risiko: tetapi juga menambah nilai melalui peningkatan dalam:

- ✓ **Peningkatan kinerja model:** Data yang bersih dan representatif menghasilkan hasil yang lebih baik.
- ✓ **Audit dan investigasi yang lebih cepat:** Dokumentasi yang baik mempercepat respons regulasi.
- ✓ **Keselarasan lintas tim:** Peran yang jelas mengurangi konflik dan pengerjaan ulang.
- ✓ **Keunggulan reputasi:** Kematangan tata kelola yang ditunjukkan membangun kepercayaan pemangku kepentingan.

Meskipun tidak selalu diwajibkan oleh hukum, tata kelola data harus dianggap sebagai keharusan strategis bagi organisasi yang mengadopsi sistem AI. Organisasi yang mengabaikan tata kelola sering mengalami lebih banyak masalah yang lebih besar, termasuk pengerjaan ulang, litigasi, dan masalah nilai merek.

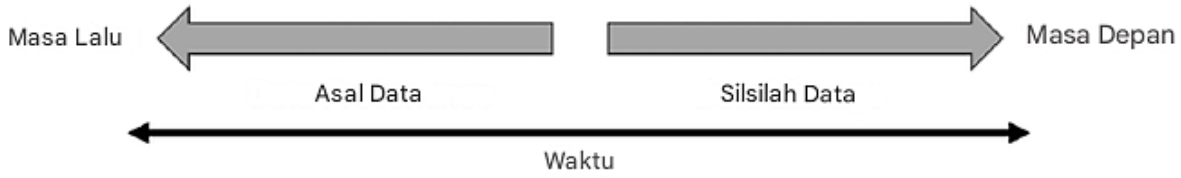
Contoh perusahaan teknologi yang mengadopsi standar tata kelola data ISO/IEC 38505-1 mengalami pengurangan waktu pelatihan model, persetujuan pemangku kepentingan yang lebih cepat, dan kepercayaan eksternal yang lebih besar pada sistem AI-nya.

9.2 SILSILAH DAN ASAL USUL DATA

Untuk membangun sistem AI yang dapat dipercaya, bertanggung jawab, dan etis, organisasi perlu memahami dari mana data berasal, bagaimana data bergerak, dan bagaimana data berubah. Konsep silsilah data dan asal usul data membentuk tulang punggung

pemahaman ini. Meskipun sebagian besar organisasi tidak mengikuti praktik ini, mendokumentasikan silsilah dan asal usul data sangat penting untuk akuntabilitas, kepatuhan terhadap peraturan, dan penjelasan model.

Tanpa pelacakan silsilah dan asal usul data yang kuat, organisasi mungkin kesulitan menjelaskan bagaimana sebuah model sampai pada outputnya, menilai keandalan data pelatihan, atau mendeteksi input yang tidak tepat atau rusak. Karena sistem AI semakin memengaruhi keputusan-keputusan penting, kebutuhan akan keterlacakan menjadi semakin mendesak. Kurangnya pelacakan ini dapat mengakibatkan bahaya. Gambar 9.1 menggambarkan asal usul data dan silsilah data.



Gambar 9.1 Asal usul data dan silsilah data.

Pelacakan silsilah dan asal usul data mungkin terasa seperti pekerjaan di balik layar (dan memang demikian), tetapi hal ini penting untuk membangun sistem AI yang kredibel, dapat dipertahankan, dan etis. Organisasi yang mengabaikan domain ini akan menghadapi kegagalan, sementara mereka yang memprioritaskannya menciptakan sistem AI yang layak dipercaya, menciptakan keunggulan kompetitif di sektor swasta.

Mendefinisikan Silsilah Data dan Asal Usul Data

Sangat penting untuk membedakan dengan jelas silsilah data dan asal usul data, karena keduanya sering digunakan secara bergantian tetapi merujuk pada konsep yang berbeda (meskipun terkait).

Silsilah data merujuk pada siklus hidup data dalam suatu sistem. Silsilah data mendokumentasikan bagaimana data mengalir melalui saluran, mulai dari penyerapan, penyimpanan, transformasi, konsumsi, pemeliharaan, dan pembuangan. Silsilah melacak "bagaimana dan di mana" data termasuk:

- Bagaimana data tersebut dihasilkan (atau dari mana data tersebut diperoleh—di sinilah silsilah data dan asal usul data bertemu)
- Di mana data tersebut disimpan
- Proses apa yang mengubahnya
- Model atau sistem mana yang menggunakannya
- Proses apa yang melakukan modifikasi, koreksi, dan penghapusan
- Proses apa yang melakukan operasi retensi data
- Bagaimana dan kapan data tersebut dibuang ketika tidak lagi dibutuhkan

Asal usul data mengacu pada asal dan karakteristik asli dari suatu dataset, termasuk:

- Dari mana data tersebut diperoleh
- Siapa yang membuat atau mengumpulkan data

- Dalam kondisi apa data tersebut dikumpulkan
- Apa tujuan penggunaannya

Misalnya, dataset perilaku pelanggan suatu organisasi yang digunakan dalam mesin rekomendasi AI mungkin memiliki silsilah yang jelas di dalam data lake perusahaan. Namun, asal usulnya mungkin tidak diketahui jika awalnya diperoleh dari pihak ketiga dengan dokumentasi yang tidak memadai. Silsilah dan asal usul.

Jika dilihat bersama-sama, silsilah dan asal usul data mendukung gambaran lengkap tentang kepercayaan data. Keduanya membantu organisasi mengidentifikasi sumber kesalahan dan bias, mereproduksi hasil (setidaknya untuk model AI yang dapat dijelaskan), dan menunjukkan kepatuhan terhadap peraturan yang berlaku dan kewajiban hukum lainnya.

Peran Silsilah dan Asal Usul dalam AI

Dalam sistem AI, konsekuensi dari pelacakan silsilah dan asal usul yang buruk dapat sangat signifikan. Karena sistem AI itu sendiri merupakan pengganda kekuatan, model berbasis data yang mereka gunakan memperkuat kualitas, asal, dan riwayat transformasi input yang mereka konsumsi. Atau, dengan kata lain, sampah masuk, sampah keluar. Atau, lebih jauh lagi, jika Anda tidak melacak asal usul dan transformasi data pelatihan sistem AI, bagaimana Anda berharap untuk memahami atau memecahkan masalah hasilnya?

- **Kemampuan menjelaskan dan kepercayaan:** Ketika pemangku kepentingan mempertanyakan output atau keputusan sistem AI, seperti mengapa seorang pelamar pekerjaan ditolak, pengembangnya harus dapat melacak data input model. Tanpa silsilah dan asal usul, penjelasan menjadi spekulatif, atau lebih buruk.
- **Keamanan dan integritas data:** Kurangnya silsilah dapat menyembunyikan perusakan data, peracunan, dan modifikasi yang tidak sah.
- **Kepatuhan dan kemampuan audit:** Banyak kerangka peraturan mengharuskan organisasi untuk mendokumentasikan praktik penanganan data mereka. Misalnya, Undang-Undang AI Uni Eropa mensyaratkan keterlacakan untuk data sumber dan transformasinya.

Mereproduksi perilaku model untuk debugging, pelatihan ulang, atau litigasi memerlukan pemahaman yang jelas tentang perjalanan dan asal usul data. Silsilah dan asal usul data mendukung reproduksibilitas dan, oleh karena itu, akuntabilitas.

Sebagai contoh, sebuah lembaga transportasi menerapkan model prediktif untuk mengoptimalkan rute bus. Enam bulan kemudian, mereka ingin melatih ulang model tersebut tetapi tidak dapat menemukan dataset asli atau merekonstruksi transformasi fiturnya. Proyek tersebut ditangguhkan hingga lembaga tersebut dapat menentukan jalan keluar.

Mencatat Asal Usul Data

Pelacakan asal usul data yang efektif dimulai dengan pengumpulan dan dokumentasi metadata. Metadata tersebut harus mencakup:

- Entitas sumber (pemilik, organisasi, sistem)
- Metode dan tanggal pengumpulan
- Informasi versi
- Lisensi data (ketentuan penggunaan)

- Hak akses
- Penggunaan yang dimaksudkan atau konteks asli
- Silsilah data
- Pertimbangan yurisdiksi (di mana dikumpulkan, peraturan yang berlaku) Konteks (penggunaan yang dimaksudkan, asumsi, batasan)

Metadata harus menjadi elemen wajib yang merupakan bagian dari paket dokumentasi keseluruhan untuk sistem AI. Tata kelola AI dan tata kelola data dapat menegakkan hal ini melalui kebijakan.

Data yang diperoleh dari pihak ketiga, termasuk vendor, broker, mitra, atau repositori publik, mungkin mengandung risiko tambahan. Kebijakan tata kelola data harus mensyaratkan bahwa organisasi memperoleh deklarasi asal usul data dari sumber data, serta bukti pengumpulan dan persetujuan yang sah, dan dokumentasi yang menjelaskan setiap modifikasi atau agregasi yang dilakukan oleh sumber tersebut.

Misalnya, pengembang sistem di suatu organisasi menggunakan dataset gambar sumber terbuka untuk melatih model pengenalan wajah. Setelah penerapan, organisasi tersebut mengetahui bahwa dataset tersebut mencakup gambar berhak cipta yang diperoleh tanpa izin. Akibatnya, organisasi tersebut sekarang menghadapi risiko hukum dan potensi kerusakan pada mereknya.

Standar, termasuk W3C PROV atau ISO/IEC 5181 (Kecerdasan Buatan—Asal Usul Data dan Sistem AI), mendefinisikan model formal untuk pelacakan asal usul. Adopsinya terbatas, dan ISO/IEC 5181 masih dalam tahap pengembangan. Meskipun demikian, standar ini dapat memberikan arahan bagi organisasi yang ingin mengadopsi pendekatan standar untuk mengumpulkan dan mendokumentasikan asal usul data.

Pelacakan Silsilah Data

Sementara asal usul data melihat ke belakang ke asal, silsilah melacak jalur ke depan dalam waktu. Catatan silsilah harus mencakup:

- Sumber dan format data
- Lokasi penyimpanan
- Riwayat versi
- Skrip transformasi atau proses ETL (ekstrak, transformasi, muat) Jalur pipeline data
- Sistem dan model yang menggunakan dataset
- Proses retensi dan pembuangan data

Silsilah mundur, yang terkadang dikacaukan dengan asal usul data, melacak setiap titik data kembali ke asalnya. Silsilah maju melacak bagaimana elemen data mengalir melalui sistem untuk memengaruhi proses hilir.

Platform pengamatan data modern, seperti Monte Carlo, Collibra, atau OpenLineage, menawarkan alat visualisasi untuk merepresentasikan silsilah secara grafis. Alat-alat ini sangat berharga ketika beberapa tim berinteraksi dengan dataset bersama.

Sebuah perusahaan contoh memperhatikan lonjakan positif palsu dari model AI, yang dilacak ke skrip transformasi yang salah format yang diperkenalkan dua minggu sebelumnya.

Alat pelacakan silsilah membantu mengisolasi perubahan, memungkinkan pengembalian (rollback) pada alur kerja yang terpengaruh.

Mode Kegagalan Umum

Banyak organisasi gagal menerapkan pelacakan silsilah dan asal usul yang efektif karena hambatan struktural dan budaya. Organisasi memutuskan bahwa silsilah tidak penting atau sama sekali tidak menyadari praktik tersebut. Beberapa mode kegagalan meliputi:

- **Silo data dan organisasi:** Silo data dan silo organisasi menghambat silsilah dan asal usul yang efektif. Ketika kumpulan data ditransfer antar departemen tanpa kepemilikan yang jelas atau standar yang ditetapkan, silsilah dan asal usul keduanya dapat menjadi tidak jelas. Kerangka kerja tata kelola harus mensyaratkan dokumentasi yang konsisten di seluruh silo.
- **Penggunaan data informal:** Pengembangan AI bayangan, di mana ilmuwan data menerapkan model AI dan memanipulasi kumpulan data di luar alur kerja yang disetujui, menciptakan potensi titik buta. Tanpa tata kelola, proyek ad hoc ini dapat memasukkan data yang tidak dapat diandalkan ke dalam model produksi yang tidak disetujui, yang menyebabkan hasil yang tidak akurat.
- **Proses manual dan terfragmentasi:** Jika dokumentasi manual atau tersebar di berbagai alat, tugas, dan prosedur, seringkali diabaikan. Otomatisasi dapat membantu menyediakan penangkapan silsilah dan asal usul yang konsisten.

Sebagai contoh, sebuah perusahaan ritel besar mendokumentasikan transformasi secara manual dalam spreadsheet. Seorang karyawan magang mengganti nama kolom fitur dan lupa memperbarui spreadsheet tersebut. Beberapa bulan kemudian, kesalahan tersebut menyebabkan prediksi inventaris yang salah selama musim liburan.

Membangun Kerangka Kerja Asal Usul dan Silsilah

Ketelusuran tidak terjadi begitu saja; organisasi perlu bersikap proaktif terhadap asal usul dan silsilah, menetapkan kebijakan, alat, dan proses untuk mengoperasionalkan ketelusuran. Menetapkan asal usul dan silsilah data dimulai dengan kebijakan, yang harus mendefinisikan:

- ✓ Bidang metadata wajib untuk dataset baru
- ✓ Kapan dan bagaimana asal usul data harus diverifikasi dan disetujui
- ✓ Tanggung jawab kepemilikan untuk dokumentasi asal usul dan silsilah data
- ✓ Persyaratan untuk pengambilan data otomatis dan jejak audit
- ✓ Grafik pelacakan yang dibutuhkan (misalnya, silsilah data tingkat dataset, tingkat record, atau tingkat fitur)

Dalam tumpukan teknologi, kerangka kerja asal usul dan silsilah data dapat mencakup:

- ➢ Repositori metadata (misalnya, katalog data)
- ➢ Visualisasi silsilah data
- ➢ Alat ETL dengan pelacakan dan pencatatan terintegrasi
- ➢ Sistem kontrol versi untuk dataset dan skrip transformasi
- ➢ Kontrol keamanan (misalnya, enkripsi dan manajemen akses) untuk melindungi catatan silsilah dan asal usul data itu sendiri

Data asal usul dan silsilah dapat diintegrasikan dengan alat MLOps, seperti MLflow, DVC (Kontrol Versi Data), atau Amazon SageMaker, yang memungkinkan ketertelusuran dari data mentah melalui pelatihan hingga penerapan.

Banyak organisasi menggunakan kartu model (pertama kali dibahas di Bab 2) untuk menjelaskan perilaku model secara detail. Kartu model dapat menyertakan referensi ke data silsilah dan asal-usul, sehingga meningkatkan transparansi bagi pemangku kepentingan dan regulator, serta menghubungkan model AI dengan kumpulan data yang sesuai.

Kasus Penggunaan dan Contoh Industri

Praktik ketertelusuran dapat bervariasi menurut industri, tetapi prinsip-prinsip ketertelusuran secara umum berlaku. Beberapa contoh meliputi:

- **Layanan keuangan:** Regulator mungkin memerlukan silsilah yang jelas untuk semua data yang digunakan dalam penilaian kredit, deteksi penipuan, dan perdagangan algoritmik. Dokumentasi harus menunjukkan bagaimana keputusan dibuat dan data apa yang memengaruhinya.
- **Perawatan kesehatan:** Kepatuhan HIPAA dan kekhawatiran tentang keamanan data pasien mengharuskan semua sistem AI terkait kesehatan dapat melacak data kembali ke sumber yang divalidasi. Asal usul data uji klinis atau rekam medis elektronik sering kali diperiksa dengan cermat.
- **Sektor publik:** Transparansi sangat penting dalam proyek AI pemerintah. Kemampuan audit sumber data memastikan akuntabilitas demokratis, khususnya dalam peradilan pidana atau otomatisasi kesejahteraan.

Contoh lembaga kota menggunakan AI untuk menandai remaja berisiko berdasarkan kinerja sekolah dan data lingkungan. Setelah kekhawatiran muncul, pejabat menggunakan silsilah yang terdokumentasi untuk menunjukkan bahwa model tersebut tidak menggunakan atribut yang dilindungi, seperti ras atau tingkat pendapatan.

Manfaat Keterlacakan yang Kuat

Organisasi yang berinvestasi dalam kemampuan silsilah dan asal usul data menuai manfaat di berbagai bidang teknis, hukum, dan bisnis, termasuk:

- Pengurangan waktu debugging dan penyelesaian masalah model yang lebih cepat untuk masalah yang terkait dengan data pelatihan
- Kepercayaan yang lebih besar pada output model bagi para pemangku kepentingan
- Peningkatan respons audit dan postur regulasi
- Kualitas data yang lebih tinggi menghasilkan lebih sedikit masalah terkait data, dan penyelesaian masalah yang lebih cepat

Tantangan dan Kompromi Keterlacakan

Membangun program asal usul dan silsilah data yang komprehensif menghadirkan hambatan teknis dan organisasi. Kompromonya terletak pada menciptakan biaya tambahan dari proses dan alat yang menyediakan data penting, atau mengabaikan keterlacakan untuk meningkatkan kelincahan.

Pelacakan silsilah data otomatis dapat menambah latensi pada pipeline data dan meningkatkan kebutuhan penyimpanan. Tim tata kelola harus bekerja sama dengan tim teknik untuk menemukan keseimbangan yang tepat antara keterlacakan dan kinerja.

Organisasi yang menggunakan model pihak ketiga atau sumber terbuka mungkin tidak dapat memperoleh data asal usul atau silsilah data lengkap, karena ini dapat dianggap sebagai kekayaan intelektual. Organisasi harus mengevaluasi keterlacakan saat memilih alat eksternal untuk memastikan kinerja dan keandalan yang optimal.

Organisasi perlu menentukan tingkat investasi yang dibutuhkan untuk membangun infrastruktur manajemen metadata, audit, dan visualisasi. Seperti inisiatif lainnya, ketertelusuran itu sendiri mungkin memerlukan pengembangan studi kelayakan bisnis. Seperti inisiatif serupa, banyak organisasi akan kekurangan sumber daya untuk fungsi-fungsi ini sampai terpaksa melakukannya karena insiden atau investigasi.

9.3 PENGUJIAN SISTEM AI

Seperti halnya sistem atau aplikasi baru atau yang diperbarui, pengujian sistem AI sangat penting untuk memastikan kualitas, keandalan, dan kepercayaan model. Tidak seperti pengujian perangkat lunak tradisional, yang berfokus pada verifikasi keluaran yang diharapkan untuk masukan tertentu, pengujian sistem AI mencakup berbagai metode pengujian yang lebih luas. Ini termasuk kinerja statistik, keadilan, keamanan, kemampuan menjelaskan, dan integrasi dalam sistem yang lebih besar.

Dan, seperti aplikasi bisnis lainnya, sistem AI juga perlu diuji secara iteratif dan terus-menerus, karena model sering berevolusi melalui pelatihan ulang, penyempurnaan, atau adaptasi terhadap data baru. Pengujian menyeluruh selama siklus pelatihan dan penerapan memverifikasi perilaku model, dan mengungkap risiko dan kelemahan yang mungkin tetap tersembunyi hingga pasca-penerapan.

Pengujian AI Adalah Sebuah Disiplin

Pengujian AI bukanlah aktivitas yang berlaku untuk semua, melainkan praktik lintas disiplin yang menggabungkan tujuan bisnis, rekayasa, statistik, etika, dan keahlian domain. Jika dilakukan dengan benar, pengujian AI membantu memastikan bahwa sistem AI tidak hanya fungsional tetapi juga adil, aman, patuh, dan selaras dengan tujuan bisnis yang dimaksudkan.

Seiring dengan terus meningkatnya adopsi AI, organisasi yang menggunakan AI harus meningkatkan praktik pengujian mereka dari validasi ad hoc menjadi proses formal, terdokumentasi, dan dapat diaudit. Pengujian bukan hanya tugas teknis tetapi juga tanggung jawab tata kelola dan pendorong kepercayaan jangka panjang.

Memahami Tujuan Pengujian AI

Pengujian sistem AI melayani berbagai tujuan, termasuk menilai akurasi, memastikan keamanan, mengevaluasi keadilan, dan membangun kepercayaan pemangku kepentingan. Pengujian AI berbeda dari pengujian aplikasi tradisional karena melibatkan perilaku stokastik, ketergantungan data, dan pergeseran batas kinerja. Pengujian sistem AI harus mencakup:

- Variabilitas dalam keluaran model
- Kinerja pada kasus yang tidak terlihat atau kasus ekstrem

- Kompromi antara kinerja dan interpretasi
- Kepatuhan terhadap standar peraturan yang terus berkembang

Pengujian model yang lolos tolok ukur akurasi mungkin masih menghasilkan keluaran yang bias atau berbahaya dalam produksi, jika kasus ekstrem atau keadilan tidak diidentifikasi atau diuji secara memadai.

Pengujian harus dilakukan sepanjang siklus hidup AI:

- Selama pelatihan (untuk memantau konvergensi dan perilaku)
- Dalam validasi (untuk memverifikasi keluaran dan keputusan)
- Pra-penyebaran (untuk menguji ketahanan terhadap masukan yang merugikan atau tidak terduga)
- Setelah penyebaran (untuk mendeteksi pergeseran data atau penurunan kinerja)

Masing-masing fase ini membutuhkan strategi dan metrik pengujian yang berbeda.

Pengujian Unit dan Komponen

Pengujian unit dalam AI berfungsi untuk memvalidasi fungsionalitas komponen individual dalam sistem secara keseluruhan, termasuk logika pemrosesan data dan fungsi model.

Transformasi data input seperti normalisasi, pengkodean, atau imputasi harus diuji kebenarannya dan konsistensinya. Bug dalam langkah-langkah ini dapat menimbulkan masalah pada model hilir. Pembungkus model (perangkat lunak yang membungkus model AI), fungsi penilaian, dan logika evaluasi harus diisolasi dan diuji, menggunakan input sintetis untuk memverifikasi penanganan yang benar terhadap kasus-kasus ekstrem (misalnya, nilai null, nilai ekstrem, data input yang salah format). Pengujian unit dapat diintegrasikan ke dalam alat dan alur kerja integrasi berkelanjutan (CI) menggunakan kerangka kerja pengujian (seperti PyTest atau unittest di Python), memastikan perilaku yang konsisten di seluruh perubahan kode.

Pengujian Integrasi

Pengujian integrasi memvalidasi interaksi antar komponen: penyerapan data, inferensi model, penyimpanan, dan sistem hilir. Setiap titik dalam sistem secara keseluruhan di mana satu komponen berkomunikasi dengan komponen lain adalah item yang harus diuji, untuk memastikan bahwa setiap elemen bekerja dengan benar dan bahwa data atau instruksi yang diteruskan di antara mereka terstruktur dengan benar.

Data input harus dibersihkan untuk memastikan bahwa data yang masuk sesuai dengan format, nilai, dan set karakter yang diharapkan, dan bahwa API atau antarmuka berfungsi sebagaimana mestinya di bawah beban. Misalnya, model yang dilatih pada usia numerik mungkin gagal ketika sistem langsung memberikan usia sebagai string (misalnya, "35 tahun"). Pengujian integrasi akan mengidentifikasi ketidaksesuaian tipe sebelum peluncuran produksi. Pengujian antarmuka sangat penting ketika antarmuka menghubungkan sistem dari berbagai jenis (misalnya, antara UNIX atau Windows dan mainframe). Pengujian integrasi memastikan bahwa prediksi model ditangani dengan benar oleh aplikasi, dasbor, atau sistem pengambilan keputusan.

Percaya tetapi Verifikasi

Pengujian integrasi harus mensimulasikan alur kerja sistem lengkap dari ujung ke ujung, termasuk kondisi kegagalan (misalnya, data yang hilang, input yang tidak terduga, atau layanan yang tidak tersedia), untuk memvalidasi ketahanan sistem dalam kondisi dunia nyata.

Pengujian Validasi

Tujuan pengujian validasi adalah untuk memastikan apakah sistem AI berkinerja sesuai yang diharapkan dalam aplikasi dunia nyata. Pengujian validasi dilakukan setelah pengujian unit dan integrasi, yang memeriksa apakah sistem berfungsi secara teknis. Pengujian validasi melangkah lebih jauh dengan membandingkan output sistem AI dengan harapan bisnis, persyaratan etika, dan standar kepatuhan.

Dengan kata lain, pengujian validasi memastikan apakah sistem AI memecahkan masalah bisnis yang dinyatakan. Seperti jenis pengujian lainnya, kriteria keberhasilan harus didefinisikan sebelum pengujian dimulai. Jika tidak, akan sulit untuk menentukan apakah pengujian validasi dianggap berhasil.

Metrik yang terkait dengan pengujian validasi meliputi akurasi, presisi, recall, dan skor F1. Misalnya, dalam sistem deteksi penipuan penerbit kartu kredit, presisi lebih penting daripada recall untuk menghindari tuduhan palsu. Pengujian validasi model AI berfokus pada pengoptimalan keseimbangan ini.

Pengujian Kinerja

Organisasi menggunakan pengujian kinerja untuk memastikan sistem AI dapat memenuhi target kecepatan, skalabilitas, dan ketersediaan di lingkungan operasionalnya. Beberapa metode tersebut meliputi:

- **Latensi inferensi:** Waktu yang dibutuhkan sistem AI untuk mengembalikan prediksi dari input dikenal sebagai latensi inferensi. Sistem waktu nyata membutuhkan inferensi dengan latensi rendah. Pengujian kinerja di bawah beban simulasi mengungkapkan hambatan dalam eksekusi model atau I/O.
- **Throughput dan skalabilitas:** Untuk tugas pemrosesan batch, organisasi menguji berapa banyak input yang dapat ditangani sistem per satuan waktu, yang dikenal sebagai throughput. Alat seperti LoadRunner, JMeter, atau Locust dapat mensimulasikan beban.
- **Pemanfaatan sumber daya:** Penting untuk mengetahui berapa banyak sumber daya CPU, GPU, memori, penyimpanan, dan jaringan yang dibutuhkan sistem AI. Organisasi sering melakukan pengujian beban rata-rata dan beban puncak, untuk memastikan sistem berjalan dalam batas infrastruktur.

Misalnya, mesin rekomendasi produk organisasi ritel gagal memberikan hasil di bawah lalu lintas liburan yang padat. Pengujian beban mengungkapkan kebutuhan akan infrastruktur yang lebih baik yang dapat beroperasi di bawah beban puncak.

Uji yang Terburuk, Bukan yang Rata-Rata
Model yang memiliki waktu inferensi rata-rata 50 ms tetapi melonjak hingga 2 detik di bawah beban tertentu masih dapat menurunkan pengalaman pengguna.

Pengujian Keamanan

Meskipun memiliki beberapa properti unik, sistem AI adalah sistem informasi, yang secara inheren rentan terhadap berbagai bentuk serangan. Selain kerentanan konvensional, model AI rentan terhadap berbagai ancaman keamanan, termasuk serangan adversarial dan inversi model.

Beberapa pengujian khusus AI yang harus dilakukan organisasi meliputi:

- ✓ **Pengujian adversarial:** Suntikkan input yang dirancang khusus untuk menyebabkan prediksi yang salah. Misalnya, menambahkan noise halus pada gambar dapat menyebabkan kesalahan klasifikasi.
- ✓ **Peracunan data dan pintu belakang:** Nilai integritas data pelatihan untuk mendeteksi input berbahaya yang dapat memengaruhi perilaku model dalam kondisi tertentu.
- ✓ **Ekstraksi dan inversi model:** Organisasi dapat menguji apakah penyerang dapat merekonstruksi data pelatihan atau merekayasa balik perilaku model melalui kueri berulang.

Organisasi juga perlu melakukan pengujian keamanan komprehensif, termasuk:

- ➢ Pemindaian kerentanan pada tingkat aplikasi, sistem, dan jaringan
- ➢ Pengujian penetrasi pada tingkat aplikasi, sistem, dan jaringan
- ➢ Pengujian keamanan rantai pasokan untuk model dan dependensi pihak ketiga
- ➢ Pengujian penetrasi fisik untuk sistem di tempat
- ➢ Pengujian rekayasa sosial untuk menguji ketahanan tenaga kerja

Contoh model pengenalan wajah diserang dengan ribuan gambar probing. Penyerang merekonstruksi wajah pelatihan dengan fidelitas tinggi. Pengujian keamanan dapat mengidentifikasi kerentanan ini, mendorong pengurangan paparan model melalui pembatasan laju atau privasi diferensial.

Pengujian Bias dan Keadilan

Seperti yang dibahas di seluruh buku ini, bias adalah kelemahan potensial yang melekat pada sistem AI apa pun, di mana outputnya secara sistematis merugikan individu atau kelompok tertentu. Beberapa jenis pengujian meliputi:

- ❖ Kesamaan kinerja: Mengukur bagaimana metrik kinerja (misalnya, akurasi, tingkat positif palsu) bervariasi di berbagai kelompok demografis. Mengidentifikasi perbedaan yang tidak beralasan yang mungkin diakibatkan oleh bias proksi.
- ❖ Keadilan kontrafaktual: Evaluasi apakah mengubah atribut yang dilindungi (misalnya, jenis kelamin, afiliasi agama, afiliasi politik) mengubah keluaran model secara tidak tepat.

Alat untuk Pengujian Bias meliputi AIF360, Fairlearn, dan What-If Tool dari TensorFlow.

Sebagai contoh, model kredit bank menolak pelamar dari kalangan minoritas dua kali lebih banyak daripada pelamar kulit putih, meskipun profil kredit mereka setara. Pengujian keadilan mengungkapkan kesenjangan tersebut dan mendorong upaya pelatihan ulang.

Risiko Hukum
Di banyak yurisdiksi, algoritma yang bias tidak hanya tidak etis tetapi juga ilegal. Pengujian keadilan membantu organisasi mengidentifikasi masalah keadilan, memungkinkan mereka untuk menghindari denda peraturan dan reaksi negatif publik.

Pengujian Interpretasi dan Penjelasan

Pengujian interpretasi digunakan untuk menentukan apakah perilaku model AI dapat dipahami secara bermakna oleh organisasi, dan karenanya dapat dijelaskan kepada pengguna, regulator, dan pengembang. Jenis pengujian meliputi:

- Metode agnostik model: Alat seperti SHAP atau LIME dapat membantu menjelaskan prediksi individual, yang dapat membantu mengidentifikasi korelasi palsu atau artefak data.
- Interpretasi global: Sangat penting untuk dapat meringkas apa yang secara umum dihargai oleh model AI. Analisis kepentingan fitur, pohon keputusan, atau inspeksi koefisien (dalam model linier) dapat memberikan wawasan.
- Kemampuan simulasi: Apakah manusia dapat mereplikasi atau memahami prediksi model dikenal sebagai kemampuan simulasi. Model yang lebih sederhana (misalnya, regresi logistik atau sistem berbasis aturan) mungkin lebih disukai dalam kasus penggunaan berisiko tinggi karena alasan ini.

Contoh sistem AI perawatan kesehatan memprediksi risiko penyakit jantung. Nilai SHAP menunjukkan bahwa kode pos adalah pendorong utama, menunjukkan potensi proksi untuk ras atau status sosial ekonomi.

Pengujian dengan Interaksi Manusia dan Skenario

Banyak organisasi menerapkan sistem AI yang berinteraksi dengan manusia secara kritis, termasuk peninjauan keputusan sistem AI atau jenis keluaran lainnya. Pengujian harus mencakup evaluasi berbasis skenario yang mensimulasikan perilaku pengguna dan jalur pengambilan keputusan di dunia nyata. Jenis pengujian meliputi:

- Alur pengguna ujung ke ujung: Pengujian ini mengevaluasi perilaku sistem saat berinteraksi dengan pengguna nyata. Pengujian ini dapat mengungkap hasil yang tidak terduga karena desain UX, miskomunikasi, atau kasus-kasus ekstrem yang tidak terduga.
- Pengujian interaktif dan umpan balik: Untuk sistem AI yang memberikan rekomendasi atau membantu pengambilan keputusan, respons pengguna dapat memengaruhi perilaku sistem di masa mendatang. Pengujian sistem harus mempertimbangkan dinamika jangka panjang dan efek umpan balik.

Contoh alat pendukung keputusan medis berbasis AI-manusia memberikan rekomendasi yang terkadang diabaikan oleh dokter. Pengujian pasca-implementasi mengungkapkan bahwa skor

kepercayaan kurang terkalibrasi dengan baik, yang berarti pengguna tidak mempercayai sistem AI.

Dokumentasi dan Keterlacakan

Semua jenis pengujian yang dibahas di bagian ini harus didokumentasikan, termasuk prosedur pengujian, rencana pengujian, dan hasilnya. Artefak pengujian ini mendukung transparansi, kemampuan audit, dan reproduksibilitas.

Pustaka dokumen setiap sistem AI harus mencakup rencana pengujian yang mendefinisikan:

- Tujuan
- Lingkup dan tanggung jawab
- Skenario pengujian
- Metrik dan ambang batas
- Kriteria lulus/gagal

Hasil semua pengujian harus didokumentasikan, termasuk:

- Hasil pengujian
- Masalah yang ditemui
- Langkah-langkah perbaikan yang diidentifikasi
- Tinjauan pasca-tindakan
- Referensi ke dataset, versi kode, dan lingkungan yang digunakan

Dibahas di tempat lain dalam buku ini, kartu model dan lembar fakta model merangkum perilaku model, dataset pelatihan/pengujian, kinerja, kasus penggunaan yang dimaksudkan, dan keterbatasan yang diketahui. Mereka juga memberikan pemangku kepentingan pemahaman tingkat tinggi tentang apa yang dirancang—dan tidak dirancang—untuk dilakukan oleh model tersebut.

Misalnya, kartu model untuk penyaring resume bertenaga AI suatu organisasi mengungkapkan bahwa kinerja menurun untuk pelamar tanpa gelar sarjana. Organisasi menandai hal ini untuk analisis dan potensi mitigasi pada rilis berikutnya.

Praktik Organisasi untuk Pengujian AI

Pengujian AI harus rutin, dapat diulang, terdokumentasi, dan tertanam dalam alur kerja proses tim, struktur akuntabilitas, dan kerangka kerja tata kelola.

Praktik penting meliputi:

- ✓ Menetapkan tanggung jawab pengujian secara jelas. Di sektor yang diatur, ini mungkin melibatkan validasi pihak ketiga atau tim risiko terpisah.
- ✓ Menetapkan gerbang tinjauan formal dalam proses pengembangan siklus hidup, di mana bukti pengujian dievaluasi sebelum model dapat dilanjutkan ke penerapan.
- ✓ Organisasi perlu menumbuhkan budaya di mana pengujian bukan hanya formalitas tetapi mekanisme kontrol kualitas dan risiko utama. Insentif, termasuk penghargaan untuk penemuan masalah dan transparansi, membantu meningkatkan hasil pengujian.

Tantangan dan Kompromi

Organisasi menemukan bahwa pengujian sistem AI menantang karena sifat probabilistik dan kompleksitas dunia nyata dari sistem ini. Tidak selalu mudah untuk merancang pengujian untuk mengatasi masalah tertentu.

Misalnya, dalam beberapa kasus penggunaan AI, tidak ada jawaban "benar" yang pasti, sehingga validasi secara inheren bersifat subjektif dan kabur. Selain itu, keadilan dapat didefinisikan dalam berbagai cara yang seringkali saling bertentangan (misalnya, kesempatan yang sama vs. hasil yang sama), sehingga mendorong organisasi untuk memutuskan prinsip mana yang harus diprioritaskan. Terakhir, beberapa model berkinerja tinggi, seperti jaringan saraf dalam (deep neural networks), mungkin kurang dapat dijelaskan, sehingga pengujian integritas dan akurasi menjadi lebih menantang. Terkadang, pengujian harus menyeimbangkan antara kinerja, kepercayaan pemangku kepentingan, dan persyaratan peraturan.

9.4 MENGIDENTIFIKASI DAN MENGELOLA RISIKO PENGUJIAN

Seperti yang dinyatakan di seluruh bab ini, sistem AI harus diuji secara menyeluruh sebelum diterapkan; namun, pengujian bukan hanya tentang memverifikasi fungsionalitas. Pengujian juga tentang mengidentifikasi risiko yang muncul selama pelatihan dan validasi; risiko ini dapat muncul dari data yang cacat, metrik yang tidak tepat, cakupan pengujian yang buruk, atau ekspektasi kinerja yang tidak sesuai.

Model AI berbeda secara signifikan dari aplikasi perangkat lunak tradisional: perilakunya dipelajari, bukan diprogram secara eksplisit. Lebih lanjut, perilaku sistem AI dipelajari tidak hanya dari kode perangkat lunaknya tetapi juga dari data pelatihannya. Faktor-faktor ini membuat pengujian lebih probabilistik dan kurang deterministik. Sistem AI yang berkinerja baik dalam situasi biasa mungkin masih menghasilkan keluaran yang berbahaya atau tidak adil dalam beberapa skenario.

Ketidakpastian ini harus diidentifikasi dan dinilai selama fase pelatihan dan pengujian pengembangan sistem AI sebagai bagian dari keseluruhan proses pengembangan sistem AI. Idealnya, aktivitas pelatihan dan pengujian ini secara khusus menyebutkan identifikasi risiko. Ada risiko teknis, etis, dan organisasi yang muncul selama tahap-tahap ini, dan semuanya harus ditangani secara proaktif.

Mengelola risiko selama pelatihan dan pengujian harus dipandang sebagai gaya hidup, bahkan bagian dari budaya organisasi, bukan hanya aktivitas sekali saja. Pengujian lebih tentang tanggung jawab daripada kinerja. Risiko mencakup aspek teknis, hukum, dan etika. Dengan mengadopsi pendekatan yang ketat, terstruktur, dan kolaboratif untuk pengujian AI, organisasi dapat membangun sistem yang tidak hanya cerdas tetapi juga aman, adil, dan dapat dipercaya.

Sifat Unik Pengujian AI

Seperti yang dinyatakan sebelumnya dalam bab ini, pengujian AI berbeda dari pengujian perangkat lunak tradisional baik dalam tujuan maupun pendekatannya. Pengujian

AI berfokus pada perilaku di berbagai distribusi daripada kebenaran deterministik, tidak seperti aplikasi perangkat lunak konvensional.

Sementara pengujian perangkat lunak aplikasi konvensional memverifikasi bahwa input spesifik menghasilkan output yang diharapkan, pengujian sistem AI mengevaluasi distribusi kinerja, seperti presisi, recall, atau keadilan di seluruh subkelompok.

Untuk banyak aplikasi AI, jawaban "benar" (yaitu, output yang diharapkan) mungkin bersifat subjektif atau tidak tersedia. Misalnya, dalam moderasi konten atau prediksi risiko, mungkin tidak ada satu label definitif. Dengan kata lain, tidak ada jaminan kebenaran mutlak dalam beberapa sistem AI.

Pelatihan itu sendiri merupakan proses dinamis; organisasi harus memantau data dan metode pelatihan dengan cermat untuk masalah konvergensi, ketidakstabilan, overfitting, dan risiko lainnya.

Mengapa "Lulus Uji Coba" Saja Tidak Cukup

Model AI mungkin berkinerja baik pada set validasi tetapi buruk di dunia nyata, karena pergeseran dataset, kasus-kasus ekstrem baru, dan teknik-teknik adversarial yang muncul. Inovasi AI diterjemahkan menjadi permukaan serangan yang semakin luas. Pengujian sistem AI harus melampaui skor akurasi untuk mempertimbangkan dimensi risiko yang muncul dan meluas.

Jenis Risiko Pengujian dalam AI

Pengujian sistem AI itu sendiri memiliki risiko yang mencakup berbagai dimensi: teknis, etis, dan operasional. Bagian ini mengkaji jenis-jenis risiko pengujian utama dan mengeksplorasi bagaimana risiko tersebut muncul.

Risiko terkait data ini terkait dengan data yang digunakan untuk melatih sistem AI:

- **Data pelatihan yang bias:** Representasi yang tidak merata menyebabkan hasil yang tidak adil atau tidak akurat.
- **Kebisingan label:** Label yang tidak konsisten atau salah mengurangi keandalan model.
- **Pencilan dan anomali:** Nilai ekstrem atau kasus langka dapat mendistorsi pelatihan atau menyebabkan perilaku model yang tidak stabil.
- **Dataset yang tidak seimbang:** Satu kelas atau kelompok kurang terwakili, menyebabkan prediksi yang menyimpang.
- **Pergeseran konsep:** Data pelatihan menjadi usang atau tidak sesuai dengan kondisi penerapan.

Misalnya, sistem AI yang dilatih pada data medis pra-pandemi gagal mengidentifikasi gejala COVID-19 yang umum, karena pergeseran konsep temporal.

Beberapa risiko terkait model dikaitkan dengan model AI yang dipilih:

- **Overfitting:** Model mempelajari kebisingan dalam set pelatihan daripada pola umum.
- **Underfitting:** Model terlalu sederhana untuk menangkap kompleksitas dunia nyata dalam konteks model AI.

- ❖ **Kurangnya ketahanan:** Model AI berkinerja buruk pada kasus-kasus ekstrem atau pada variasi input kecil.
- ❖ **Kalibrasi yang buruk:** Skor kepercayaan model tidak sesuai dengan probabilitas sebenarnya (misalnya, model "90% yakin" tetapi hanya benar 60% dari waktu).

Risiko evaluasi terkait dengan evaluasi sistem AI:

- ✓ **Metrik yang tidak tepat:** Ketika metrik yang salah dipilih untuk mengevaluasi sistem AI, akan sulit untuk menafsirkan hasil pengujian secara akurat.
- ✓ **Kebocoran set pengujian:** Tumpang tindih antara data pelatihan dan pengujian menyebabkan peningkatan kinerja.
- ✓ **Cakupan yang tidak memadai:** Data pengujian mungkin tidak mencakup subkelompok penting atau kasus penggunaan yang mungkin terjadi di lingkungan produksi.

Risiko penyebaran dan integrasi terkait dengan perbedaan antara lingkungan pengujian dan produksi:

- ➢ **Ketidaksesuaian lingkungan:** Asumsi pelatihan tidak berlaku di lingkungan produksi karena perbedaan data, sistem operasi, dan perangkat lunak.
- ➢ **Penggunaan AI bayangan:** Model yang belum diverifikasi digunakan tanpa pengujian formal, dokumentasi, atau persetujuan.
- ➢ **Lingkaran umpan balik:** Prediksi model AI memengaruhi pengumpulan data di masa mendatang, memperkuat bias.
- ➢ **Perbedaan keamanan dan akses:** Lingkungan produksi sering memperkenalkan permukaan serangan baru (misalnya, API yang terbuka) yang tidak ada dalam pengujian.

Teknik Identifikasi Risiko Pengujian

Mengidentifikasi risiko pengujian AI membutuhkan analisis kuantitatif dan proses tinjauan terstruktur. Risiko ini meliputi:

- ❖ **Analisis Data Eksplorasi (EDA):** EDA mengungkap masalah distribusi data, ketidakseimbangan, nilai yang hilang, dan potensi outlier. Ini sering kali merupakan langkah pertama dalam mengidentifikasi risiko sebelum pelatihan model AI dimulai.
- ❖ **Audit keadilan dan bias:** Audit ini memeriksa kinerja di berbagai subkelompok (misalnya, jenis kelamin, usia, atau ras). Perbedaan kinerja dapat mengungkapkan bias sistemik yang sebelumnya tidak terlihat.
- ❖ **Analisis kesalahan:** Tinjauan manual terhadap output yang salah klasifikasi dapat mengungkapkan pola kegagalan sistematis. Pendekatan ini bermanfaat untuk aplikasi berdampak tinggi, seperti perawatan kesehatan atau peradilan pidana.

Sebagai contoh, sebuah perusahaan jasa keuangan melakukan EDA (Analisis Data Eksplorasi) dan menemukan bahwa 80% data pinjaman berasal dari kode pos perkotaan, sementara daerah pedesaan kurang terwakili. Wawasan ini mendorong mereka untuk mengumpulkan data yang lebih seimbang sebelum melatih model AI mereka.

Strategi Mitigasi Risiko

Setelah risiko diidentifikasi, organisasi harus menerapkan intervensi yang ditargetkan dan disengaja untuk mengurangi risiko tersebut. Beberapa jenis mitigasi umum dibahas di sini.

Tiga teknik untuk meningkatkan kualitas dan keragaman data adalah sebagai berikut:

- **Augmentasi data:** Menghasilkan data pelatihan tambahan untuk mengisi celah.
- **Resampling:** Menyeimbangkan distribusi kelas menggunakan upsampling atau downsampling.
- **Data sintetis:** Menggunakan data buatan yang dibuat dengan cermat untuk meningkatkan keterwakilan. Saat menggunakan data sintetis, pastikan data tersebut tidak dapat diidentifikasi kembali sebagai subjek data nyata.

Misalnya, platform perekrutan suatu organisasi menambah datasetnya dengan profil pencari kerja yang kurang terwakili, untuk mengurangi bias gender dalam mesin rekomendasi berbasis AI-nya.

Dalam situasi di mana model AI mengalami overfitting, teknik seperti regularisasi L1 (menambahkan penalti berdasarkan nilai absolut parameter model, sehingga menyederhanakan model) atau regularisasi L2 (menambahkan penalti berdasarkan kuadrat parameter model, sehingga meningkatkan stabilitas) dapat digunakan. Kedua teknik tersebut mencegah model untuk menyesuaikan noise atau detail yang tidak relevan dalam data pelatihan. Dropout adalah teknik serupa yang digunakan dalam jaringan saraf. Menggunakan model ensemble atau melakukan penumpukan model dapat mengurangi kelemahan dari pendekatan tunggal dan meningkatkan kinerja serta ketahanan secara keseluruhan.

Akurasi Saja Tidak Selalu Cukup

Dalam sistem diagnostik medis, akurasi 95% mungkin terdengar mengesankan, tetapi jika prevalensi penyakit hanya 5%, ini bisa berarti model tersebut mengklasifikasikan hampir setiap kasus sebagai "sehat." Presisi dan recall sangat penting dalam perawatan kesehatan.

Mengelola Risiko Sepanjang Siklus Pelatihan

Pengujian tidak berakhir ketika sistem AI diterapkan. Sebaliknya, pengujian terjadi sepanjang siklus pelatihan model AI, dan bahkan setelah pelatihan selesai. Meskipun sistem AI dapat dijelaskan, sistem tersebut tidak selalu dapat diprediksi. Oleh karena itu, setiap kali ada perubahan dalam parameter operasi atau perubahan dalam data pelatihan, sistem AI harus diuji regresi untuk melihat apakah perilakunya masih berada dalam batas yang dapat diterima. Jenis pengujian yang harus dilakukan diuraikan di sini.

Fase Pelatihan Awal

- Memantau perilaku konvergensi
- Mendeteksi gradien yang menghilang atau laju pembelajaran yang buruk
- Memvalidasi keputusan arsitektur model

Pemantauan Pertengahan Pelatihan

- Memvalidasi kurva kerugian
- Melacak metrik kinerja pada set validasi
- Memperhatikan ketidakstabilan atau lonjakan metrik yang tiba-tiba

Penyetelan Tahap Akhir

- Melakukan optimasi hyperparameter akhir

- Menggunakan set validasi skala besar
- Melakukan pengujian berbasis skenario atau adversarial

Pengujian Pasca-Pelatihan

- ✓ Menggunakan simulasi dunia nyata jika memungkinkan
- ✓ Mengevaluasi kemampuan penjelasan dan perilaku kontrafaktual
- ✓ Log audit dan dokumentasi untuk transparansi

Dalam contoh perusahaan pengiriman dan logistik, selama pengujian pasca-pelatihan, sistem AI untuk optimasi rantai pasokan dimulai, di mana semua barang diproses melalui satu hub. Pengujian skenario mengungkapkan jalan pintas biaya yang dipelajari yang mengabaikan kerapuhan sistem, yang diidentifikasi tepat waktu.

Alat dan Infrastruktur Pengujian

Organisasi mendapat manfaat dari alat khusus untuk mendukung pengujian dan pelacakan risiko untuk sistem AI mereka. Untuk integrasi MLOps, alat-alat seperti MLflow, SageMaker, dan Vertex AI menawarkan alur kerja untuk mengelola pelatihan, pengujian, dan penerapan dengan kemampuan audit. Platform pengujian, termasuk *Great Expectations* (GX), Checklist (di checklist.gg), dan *TensorFlow Model Analysis* (TFMA), memfasilitasi validasi data, pengujian di seluruh dimensi tugas, dan evaluasi kinerja.

Alat pemantauan produksi, seperti Arize dan Fiddler, memungkinkan pengujian pasca-penerapan dengan membandingkan hasil dunia nyata dengan hasil yang diharapkan. Pemantauan pasca-penerapan merupakan perluasan dari pengujian dalam sistem AI. Pengujian pasca-penerapan mengidentifikasi risiko baru dan memasukkannya kembali ke dalam siklus peningkatan model. Organisasi yang tidak melakukan pemantauan pasca-penerapan berisiko mengalami pergeseran model dan masalah lainnya.

Kesalahan Umum dalam Pengujian

Bahkan tim pengembangan AI yang berpengalaman dan matang pun dapat mengalami masalah selama pengujian dan manajemen risiko. Contohnya meliputi:

- ➢ **Ketergantungan berlebihan pada akurasi pengujian:** Akurasi saja dapat menutupi banyak masalah, termasuk positif palsu, negatif palsu, dan perlakuan tidak adil terhadap subkelompok.
- ➢ **Kurangnya skenario dunia nyata:** Meskipun pengujian sintetis bermanfaat, konteks dunia nyata sangat penting. Kasus-kasus ekstrem dan variabilitas lingkungan harus disertakan dalam pengujian.
- ➢ **Tanggung jawab yang terkotak-kotak:** Ketika ilmuwan data menguji model mereka sendiri tanpa tinjauan eksternal, bias dan titik buta tidak terkendali. Praktik tinjauan sejawat sebagai teknik pengembangan perangkat lunak yang penting berlaku untuk sistem AI, dan untuk alasan yang sama: pengembang terkadang tidak mengenali kesalahan mereka sendiri.
- ➢ **Ketidaksesuaian metrik dengan tujuan bisnis:** Sebuah model mungkin mengoptimalkan metrik teknis (misalnya, skor F1) tetapi gagal memberikan nilai bisnis nyata, seperti mengurangi tingkat churn pelanggan atau meningkatkan efisiensi.

- **Mengabaikan kemampuan menjelaskan:** Kinerja tinggi tidak ada artinya jika pemangku kepentingan tidak dapat memahami atau mempercayai perilaku model. Pengujian kemampuan menjelaskan (misalnya, dengan SHAP atau LIME) harus disertakan.

Studi Kasus dan Contoh

Bagian ini membahas beberapa contoh tambahan risiko pengujian sistem AI.

Bias Algoritma Perekrutan

Sebuah perusahaan teknologi melatih model perekrutan yang lebih menyukai resume yang menggunakan bahasa yang cenderung maskulin. Analisis post hoc mengungkapkan bahwa embedding kata yang digunakan selama pelatihan memperkenalkan bias halus. Perusahaan tersebut melatih ulang dengan embedding yang telah dihilangkan biasnya dan memperbarui metrik evaluasinya.

Simulasi Kendaraan Otonom

Sebuah perusahaan AV menemukan bahwa kendaraannya berkinerja buruk di terowongan, sebuah skenario yang tidak termasuk dalam set pengujian. Mereka membangun simulator mengemudi terowongan sintetis untuk mengatasi kesenjangan tersebut dan meningkatkan penanganan kasus ekstrem.

Penyimpangan di Layanan Keuangan

Sistem deteksi penipuan mulai memicu alarm palsu pada transaksi besar tetapi sah. Pengujian lebih lanjut mengungkapkan bahwa data bervolume tinggi kurang terwakili dalam pelatihan. Model diperbarui dengan pengambilan sampel bertingkat untuk menyelesaikan masalah tersebut.

9.5 RINGKASAN

Mengembangkan sistem AI yang bertanggung jawab, efektif, dan dapat dipercaya dimulai dengan mengelola data yang digunakan untuk melatihnya. Tata kelola data memainkan peran mendasar dengan memastikan bahwa penggunaan data sesuai, sah, akurat, memadai, dan sesuai dengan tujuan. Tanpa struktur tata kelola yang memadai, organisasi akan menghadapi risiko signifikan, mulai dari model yang bias dan kegagalan kepatuhan hingga terkikisnya kepercayaan pemangku kepentingan dan sanksi regulator. Meskipun tata kelola tidak selalu diwajibkan oleh hukum, hal ini secara luas dianggap sebagai prasyarat untuk penerapan AI yang terukur dan berkualitas tinggi. Praktik tata kelola yang baik mencakup penggunaan data yang sah, kontrol integritas, dan penetapan peran seperti pengelola dan pemilik data untuk memastikan akuntabilitas di seluruh siklus hidup AI.

Dokumentasi silsilah dan asal usul data sangat penting. Silsilah melacak aliran data melalui sistem, sementara asal usul berfokus pada asal dan konteks pengumpulannya. Disiplin ilmu ini mendukung ketertelusuran, penjelasan, dan reproduksibilitas dalam pengembangan AI, memungkinkan organisasi untuk memahami bagaimana data telah ditransformasikan, dimanfaatkan, atau digunakan kembali, dan apakah penggunaan tersebut tetap sesuai. Seringkali kurang memiliki kemampuan tata kelola data, banyak organisasi kesulitan menerapkan kerangka kerja ketertelusuran, terutama saat bekerja dengan data pihak ketiga

atau data historis. Tanpa silsilah dan asal usul yang jelas, organisasi mungkin kesulitan untuk mempertahankan output model mereka, memenuhi persyaratan peraturan, atau menentukan akar penyebab masalah kinerja.

Ketika tata kelola dan ketertelusuran data telah diterapkan, tantangan lain adalah pengujian yang ketat terhadap sistem AI itu sendiri. Pengujian bukanlah peristiwa tunggal tetapi proses terstruktur yang mencakup pengujian unit, pengujian integrasi, validasi, dan evaluasi kinerja sepanjang masa pakai sistem AI. Tidak seperti perangkat lunak tradisional, sistem AI memerlukan pengujian statistik dan perilaku dalam berbagai kondisi. Pengembang harus memastikan bahwa model AI berkinerja baik pada tolok ukur standar dan juga bahwa model tersebut aman, dapat diinterpretasikan, dan adil. Infrastruktur pengujian harus mencakup metrik yang tepat, simulasi dunia nyata, dan alat untuk evaluasi berkelanjutan seiring berkembangnya model dan data pelatihan.

Berkaitan erat dengan pengujian model AI adalah identifikasi dan pengelolaan risiko yang muncul selama pelatihan dan pengujian. Risiko dapat berasal dari masalah kualitas data, label yang bias, metrik evaluasi yang tidak tepat, atau kegagalan untuk menguji di seluruh subkelompok demografis. Manajemen risiko yang efektif membutuhkan budaya transparansi, gerbang tinjauan terstruktur, dan penggunaan alat MLOps untuk mendeteksi dan memperbaiki masalah sebelum model diterapkan ke produksi. Identifikasi risiko tidak boleh menjadi proses pasif; sebaliknya, personel harus memeriksa dan meninjau data dan hasil pengujian untuk secara proaktif mengidentifikasi risiko. Meskipun beberapa risiko dapat dimitigasi melalui cara teknis, yang lain membutuhkan kesadaran organisasi dan kolaborasi lintas fungsi. Beberapa mungkin memerlukan desain ulang model.

BAB 10
PENYEBARAN DAN PEMANTAUAN

10.1 KESIAPAN PRA-RILIS

Mempersiapkan sistem AI untuk penerapan produksi memerlukan dokumentasi terstruktur, kesesuaian regulasi, tolok ukur kinerja, keputusan arsitektur, dan perencanaan operasional. Di sini, organisasi beralih dari eksperimen ke eksekusi, memastikan model tersebut layak produksi, terkelola dengan baik, dan dirancang secara bertanggung jawab dan etis. Bagian ini membahas kriteria kesiapan, model penerapan, dan strategi optimasi yang memastikan rilis yang sukses.

Menilai Kesiapan Sistem AI untuk Produksi

Sebelum sistem AI dapat dipromosikan ke produksi, organisasi harus melakukan evaluasi kesiapan untuk memastikan keamanan, akuntabilitas, dan kesesuaiannya untuk konteks yang dimaksud. Evaluasi ini dimulai dengan dokumentasi yang kuat, termasuk pembuatan kartu model.

Awalnya dibahas di Bab 2, kartu model menyediakan dokumentasi terstruktur tentang penggunaan yang dimaksudkan dari model AI, data pelatihan, metrik evaluasi, keterbatasan yang diketahui, dan pertimbangan etis. Membuat kartu model mendorong tim pengembang untuk berpikir kritis tentang asumsi, keterbatasan, dan kinerja model dalam pengaturan dunia nyata. Kartu model banyak direkomendasikan dalam praktik tata kelola AI, dan selaras dengan prinsip-prinsip dalam kerangka kerja seperti NIST AI RMF dan ISO/IEC 42001.

Banyak yurisdiksi memberlakukan kewajiban hukum pada sistem AI sebelum dirilis ke produksi. Meskipun kewajiban ini bervariasi, kewajiban tersebut dapat mencakup:

- **Penilaian kesesuaian:** Diwajibkan berdasarkan Undang-Undang AI Uni Eropa untuk sistem berisiko tinggi, termasuk kontrol kualitas terdokumentasi, penilaian risiko, dan ketentuan pengawasan manusia. Persyaratan akan bervariasi di yurisdiksi lain, meskipun ini adalah praktik yang baik.
- **Audit pihak ketiga atau validasi internal:** Dewan tata kelola internal mungkin memerlukan validasi internal atau eksternal sebelum menyetujui penggunaan produksi.
- **Keterpastian asal usul data dan lisensi:** Memastikan semua dataset pelatihan dan penyempurnaan bersumber secara sah, dilisensikan dengan tepat, dan, jika data pribadi terlibat, diproses sesuai dengan peraturan privasi yang relevan (misalnya, GDPR).
- **Persyaratan yang diberlakukan oleh asuransi:** Polis asuransi siber mungkin mengharuskan organisasi yang diasuransikan untuk menerapkan berbagai kontrol dan pengamanan terkait aplikasi, seperti manajemen perubahan, tinjauan kode, tinjauan gerbang, dan penilaian risiko berkala.

Para profesional tata kelola AI harus memahami persyaratan regional dan sektor tertentu, terutama saat menerapkan sistem yang berinteraksi dengan publik, membuat keputusan penting, atau menggunakan data pribadi. Perjanjian hukum juga dapat mencakup beberapa atau semua persyaratan yang sama ini.

Kesiapan Operasional dan Evaluasi Risiko

Sebelum penerapan, model harus menjalani pengujian stres di bawah beban kerja yang diharapkan dan kasus ekstrem. Pengujian ini meliputi:

- Benchmark latensi dan throughput (khususnya untuk kasus penggunaan waktu nyata)
- Penggunaan sumber daya (memori, komputasi) di bawah beban
- Perilaku failover selama pemadaman atau penurunan kinerja
- Evaluasi terhadap perintah yang bersifat antagonis dan berbahaya

Pengujian pra-deployment ini memastikan bahwa sistem AI akan berperilaku dapat diprediksi dan andal dalam produksi, baik dalam kondisi ideal maupun abnormal atau bermusuhan. Selain kesiapan teknis, tata kelola AI harus mensyaratkan tinjauan formal terhadap potensi kerugian dan manfaat sistem. Hal ini dapat berupa:

- **Penilaian dampak AI (AIIA):** Ini mengevaluasi dampak sosial, etis, dan operasional dari model tersebut. Bab 8 membahasnya secara mendalam.
- **Penilaian risiko:** Tim harus mengartikulasikan potensi peristiwa yang merugikan (misalnya, hasil diskriminatif atau penyebaran informasi yang salah) dan menentukan strategi mitigasi yang sesuai. Hal ini dibahas dalam Bab 7 dan 8.

Pertimbangan Deployment Model AI

Ada beberapa cara untuk mengimplementasikan sistem AI. Keputusan model deployment memengaruhi kinerja, latensi, kontrol, dan postur regulasi. Setiap opsi memiliki kelebihan dan kekurangan, seperti yang ditunjukkan pada Tabel 10.1.

Tabel 10.1
Perbandingan Opsi Penerapan Sistem AI

Penyebaran Mode	Kelebihan	Kontra
Awan	Dapat diskalakan, hemat biaya, mudah diintegrasikan dengan API.	Kontrol yang lebih sedikit atas lokasi data, potensi masalah regulasi.
Lokal	Kontrol tinggi, baik untuk data sensitif, memenuhi kepatuhan yang ketat.	Infrastruktur mahal, skalabilitas lebih rendah
Tepian	Latensi sangat rendah, tidak bergantung pada koneksi jaringan, privasi pengguna.	Keterbatasan daya komputasi, kendala ukuran model
Hibrida	Menyeimbangkan kinerja dan kepatuhan dengan menggabungkan inferensi edge/on-premise dengan pelatihan atau pembaruan berbasis cloud.	Kompleksitas arsitektur bertambah, biaya integrasi lebih tinggi.

Arsitektur penerapan yang dipilih harus selaras dengan kebutuhan bisnis, harapan pengguna, persyaratan kepatuhan, dan kendala teknis. Secara pragmatis, model penerapan seharusnya telah diputuskan pada fase desain proyek sistem AI.

AI yang siap produksi mungkin memerlukan kontainerisasi (misalnya, menggunakan Kubernetes atau Docker) untuk memungkinkan konsistensi di seluruh lingkungan. Alat kontainerisasi memungkinkan penskalaan otomatis dan orkestrasi penerapan di lingkungan multi-cloud atau hibrida. Misalnya, asisten rumah pintar menjalankan inferensi di edge untuk meningkatkan responsivitas dan melindungi privasi pengguna, tetapi model bahasanya diperbarui secara berkala melalui proses sinkronisasi cloud.

Optimasi Model AI untuk Kesesuaian Produksi

Organisasi perlu memutuskan apakah akan menerapkan model AI dalam keadaan aslinya, menyempurnakannya untuk tugas-tugas spesifik, atau melengkapinya dengan mekanisme pengambilan data. Setiap jalur memiliki implikasi yang berbeda, seperti yang ditunjukkan pada Tabel 10.2.

Tabel 10.2

Opsi Adaptasi Model AI

Mendekati	**Keterangan**	**Contoh Kasus Penggunaan**
Dengan adanya	Gunakan model dasar tanpa modifikasi	Chatbot serbaguna atau alat peringkasan
Diselesaikan dengan baik	Sesuaikan model dengan data spesifik domain.	Pengklasifikasi dokumen hukum untuk ketentuan kontrak
Generasi augmentasi pengambilan (RAG)	Gunakan basis pengetahuan eksternal selama inferensi.	Asisten pencarian yang mengutip dokumen kebijakan internal.

Penyempurnaan (*fine-tuning*) meningkatkan kinerja di domain tertentu tetapi juga dapat meningkatkan biaya pemeliharaan dan kebutuhan pelatihan ulang. Strategi RAG mengurangi risiko halusinasi dan meningkatkan landasan faktual, terutama dalam lingkungan perusahaan. Opsi ini dapat dipertimbangkan pada fase desain.

Saat diterapkan pada perangkat edge atau perangkat dengan keterbatasan sumber daya, model seringkali harus dikuantisasi (misalnya, mengurangi presisi matematis dari floating point 32-bit menjadi integer 8-bit) atau didistilasi (melatih model "siswa" yang lebih kecil untuk mendekati perilaku model "guru" yang lebih besar) menjadi model yang lebih kecil. Teknik ini dapat mengurangi ukuran model dan latensi inferensi tetapi terkadang dengan mengorbankan sedikit pengurangan akurasi. Misalnya, model visi yang diterapkan pada drone menggunakan kuantisasi untuk mengurangi konsumsi daya, meskipun dengan sedikit mengorbankan kepercayaan deteksi objek.

Dalam contoh lain, sebuah perusahaan jasa keuangan menggunakan chatbot berbasis RAG untuk membantu karyawan dengan pertanyaan kepatuhan peraturan. Model AI menanyakan dokumen kebijakan internal secara real-time dan menghasilkan respons yang berdasar, mengurangi risiko hukum dari jawaban yang dihalusinasi atau usang.

Protokol Keterlibatan Manusia dan Rilis

Beberapa sistem AI, khususnya yang termasuk dalam klasifikasi risiko tinggi, harus menyertakan tinjauan manusia atau jenis kemampuan pengesampingan berpusat pada manusia lainnya. Ini mungkin melibatkan:

- **Keterlibatan Manusia (HITL):** Manusia harus menyetujui atau mengkonfirmasi sebagian atau seluruh keputusan model (misalnya, untuk aplikasi pinjaman atau diagnosis medis).
- **Keterlibatan Manusia di Balik Sistem (HOTL):** Manusia memantau output dan dapat campur tangan jika diperlukan (misalnya, dalam sistem moderasi konten).
- **Manusia sebagai Komandan:** Keputusan strategis, seperti respons insiden, penonaktifan sistem, atau protokol eskalasi, tetap berada di tangan manusia.

Tim tata kelola AI harus mendefinisikan dan menguji kontrol ini sebelum peluncuran, untuk memastikan kontrol tersebut praktis, terukur, dan efektif. Pihak-pihak kunci harus menyadari tanggung jawab mereka dan menerima pelatihan yang tepat.

Perencanaan pra-rilis tidak hanya membutuhkan prosedur rilis tetapi juga prosedur verifikasi yang mengkonfirmasi keberhasilan implementasi model AI. Selain itu, prosedur pengembalian (rollback) harus dikembangkan, jika rilis tidak berhasil. Praktik-praktik penting meliputi:

- Prosedur rilis, prosedur verifikasi, dan prosedur pengembalian yang diajukan melalui manajemen perubahan untuk ditinjau dan disetujui
- Peluncuran bertahap dengan fitur-fitur unggulan
- Penyebaran uji coba (canary deployment) untuk pengujian kinerja
- Pemantauan KPI dan indikator risiko utama
- Saluran komunikasi yang jelas untuk pelaporan status dan eskalasi masalah

Misalnya, sebuah perusahaan asuransi merilis model penjaminan (underwriting) kepada 5% pengguna terlebih dahulu, memantau perbedaan tingkat persetujuan polis sebelum memperluas rilis ke lebih banyak pengguna.

Tinjauan Gerbang Akhir dan Persetujuan Pemangku Kepentingan

Sebelum sistem AI disetujui untuk dirilis ke produksi, sistem tersebut harus melewati satu atau lebih tinjauan gerbang akhir, yang biasanya dilakukan oleh tim tata kelola multidisiplin. Tinjauan ini memastikan bahwa:

- ✓ Semua dokumentasi (termasuk kartu model, artefak desain, log audit, catatan lisensi, dan laporan pengujian) lengkap
- ✓ Mitigasi risiko diimplementasikan, atau penundaan disetujui
- ✓ Prosedur pengerasan keamanan dan respons insiden divalidasi
- ✓ Kesiapan operasional dan rencana pemantauan telah tersedia
- ✓ Kepatuhan terhadap peraturan dan kebijakan diverifikasi
- ✓ Kontrol versi dan keterlacakan model, data, dan konfigurasi dipertahankan
- ✓ Semua pengguna dan personel operasional dilatih

Langkah ini berpuncak pada persetujuan pemangku kepentingan, di mana tata kelola AI, hukum, teknik, dan pemimpin bisnis secara resmi menyetujui rilis sistem. Contoh peran untuk tinjauan gerbang akhir meliputi:

- **Hukum:** Penggunaan data, lisensi, dan kontrak
- **Teknik:** Validasi teknis dan skalabilitas
- **Bisnis:** Keselarasan nilai dan kesiapan untuk pelanggan
- **Keamanan siber:** Kerentanan dan risiko ditangani
- **Privasi:** Tinjauan penggunaan data
- **Tata Kelola:** Keadilan, akuntabilitas, dan transparansi

10.2 PERTIMBANGAN IMPLEMENTASI

Implementasi bukan hanya tonggak teknis tetapi juga transisi tata kelola. Implementasi mengalihkan tanggung jawab dari tim pengembangan ke pemangku kepentingan operasional, yang membutuhkan kerangka kerja yang lebih luas yang mencakup manajemen risiko, akuntabilitas etis, pemberdayaan pengguna, dan pemantauan serta pengawasan berkelanjutan. Bagian ini mengkaji kebijakan, prosedur, dan praktik penting yang mengatur implementasi AI, untuk memastikan sistem AI beroperasi secara bertanggung jawab, aman, dan selaras dengan tujuan yang dimaksudkan.

Penegakan Kebijakan dan Prosedur Implementasi

Protokol implementasi yang terdefinisi dengan baik menjembatani kesenjangan antara pengembangan dan operasi. Protokol ini memastikan bahwa implementasi dilakukan secara teratur, dapat diulang, dan terkelola. Komponen kunci dari kebijakan implementasi yang efektif meliputi:

- **Gerbang persetujuan:** Implementasi harus bergantung pada persetujuan sebelumnya (misalnya, dari kepatuhan, hukum, dan keamanan).
- **Prosedur rilis yang ketat:** Pengembang sistem AI tidak boleh memiliki akses administratif ke lingkungan produksi, bahkan untuk implementasi. Sebaliknya, pengembang harus mengemas sistem AI dan menyerahkannya kepada personel operasional, yang kemudian memasang komponen sistem AI sesuai dengan prosedur instalasi yang disediakan.
- **Kontrol lingkungan:** Kebijakan harus mewajibkan isolasi antara lingkungan pengembangan, staging, dan produksi, mencegah kode yang belum diverifikasi mencapai pengguna akhir, serta menjaga pengembang agar tidak masuk ke lingkungan produksi.
- **Prosedur pengembalian:** Setiap rencana penerapan harus mencakup prosedur verifikasi yang ditentukan, langkah-langkah pengembalian, kontak eskalasi, dan prosedur pengamanan kegagalan.

Contoh lembaga keuangan menggunakan daftar periksa "lanjut/tidak lanjut" internal untuk memastikan bahwa setiap model AI yang memasuki produksi telah melewati pengujian bias, verifikasi keamanan siber, dan persetujuan eksekutif.

Penerapan harus dapat dilacak, diulang, dan dibalik. Alat pengembangan seperti Git, pipeline CI/CD, dan pencatatan otomatis membantu memastikan bahwa setiap perubahan dicatat, dilacak oleh sistem kontrol versi, dapat diaudit, dan dapat dibalik. Registri model (yang menyerupai basis data manajemen konfigurasi) melacak versi, silsilah data pelatihan, dan metadata, mendukung reproduksibilitas dan tata kelola.

Apa saja isi Registri Model?
✚ Versi dan hash model ✚ Pengidentifikasi dataset pelatihan ✚ Hasil evaluasi ✚ Status penerapan (misalnya, staging, live) ✚ Kepemilikan dan informasi kontak
Praktik penerapan yang transparan mendukung pemecahan masalah, respons insiden, pembelaan hukum, dan auditabilitas.

Tata Kelola Data dalam Fase Implementasi

Model AI sering kali diimplementasikan bersamaan dengan pipeline aplikasi atau feed data yang memberikan data secara real-time. Organisasi harus menerapkan kebijakan tata kelola data pada data input (yang digunakan pada saat inferensi) dan data apa pun yang dihasilkan selama operasi AI. Aktivitas tata kelola data utama meliputi:

- **Minimisasi data:** Batasi pengumpulan data hanya pada apa yang diperlukan untuk tugas tersebut.
- **Pembatasan tujuan:** Gunakan data hanya untuk tujuan bisnis yang telah ditentukan secara formal.
- **Kebijakan retensi:** Tentukan durasi penyimpanan berbagai dataset, khususnya data pengguna atau perilaku yang dikumpulkan selama pemrosesan sistem AI.
- **Kontrol akses:** Lindungi input data real-time dengan izin akses yang terperinci.
- **Pemantauan kualitas data:** Validasi input dan deteksi anomali untuk mencegah kesalahan berantai dalam produksi.

Implementasi sering kali memicu kewajiban hukum, termasuk ketentuan aliran data lintas batas (sistem yang diimplementasikan secara global tunduk pada hukum seperti GDPR atau LGPD, yang mengatur transfer data internasional), serta persyaratan persetujuan dan pemberitahuan.

Tim implementasi dan tata kelola AI harus bekerja sama erat dengan penasihat hukum untuk memastikan bahwa praktik penanganan data tetap sesuai ketika sistem AI mulai beroperasi. Tata kelola data dibahas secara mendalam di Bab 9.

Manajemen Risiko Selama Implementasi

Implementasi sistem AI memperkenalkan risiko baru yang melampaui kinerja model AI, yang meliputi:

- **Risiko infrastruktur:** Kegagalan jaringan, gangguan cloud, dan kehabisan sumber daya dapat mengganggu sistem produksi.
- **Risiko keamanan:** Sistem AI dapat menjadi target serangan, membuatnya rentan terhadap injeksi cepat, inversi model, eksfiltrasi data, dan serangan penolakan layanan (DoS).
- **Risiko kepatuhan dan hukum:** Pelanggaran privasi, kekayaan intelektual, atau peraturan khusus sektor.
- **Risiko reputasi:** Hasil yang cacat dan kegagalan etika dapat merusak kepercayaan pengguna dan kredibilitas merek.

Seperti yang dilakukan sebelumnya dalam siklus pengembangan, tim tata kelola AI dan keamanan siber harus menginventarisasi risiko-risiko ini dan memprosesnya menggunakan penanganan risiko untuk menentukan respons risiko yang tepat (salah satu dari: menerima, mengurangi, mentransfer, atau menghindari). Mitigasi dapat melibatkan rencana keberlangsungan bisnis, pengujian tim merah (*red teaming*), dan pengujian proaktif.

Sebagai contoh, alat AI perawatan kesehatan menjalani pengujian adversarial sebelum dirilis untuk memastikan alat tersebut tahan terhadap input yang dapat menyebabkan saran perawatan yang tidak aman.

Mitigasi risiko penerapan mengharuskan organisasi untuk memantau perilaku AI setelah dirilis. Teknik pemantauan ini meliputi:

- **Deteksi penyimpangan:** Memberi peringatan ketika input atau prediksi menyimpang secara signifikan dari data pelatihan atau tren historis.
- **Deteksi bias:** Mengungkap ketidakadilan yang muncul saat sistem berinteraksi dengan populasi dunia nyata.
- **Deteksi penyalahgunaan:** Memantau penggunaan sistem AI yang tidak sah atau tidak disengaja.
- **Pemantauan penjelasan:** Memastikan penjelasan yang dihasilkan tetap akurat, konsisten, dan mudah dipahami oleh pengguna akhir.

Dapat dikatakan, teknik pemantauan lain, seperti pemantauan keamanan siber dan infrastruktur TI, juga harus diterapkan. Teknik-teknik ini meliputi pemindaian kerentanan, pengujian penetrasi, pengumpulan dan peringatan log audit, dan analisis intelijen ancaman. Organisasi yang menetapkan dasbor, ambang batas peringatan, dan alur kerja eskalasi memastikan bahwa tata kelola AI berlanjut setelah peluncuran.

Penyimpangan dalam Penerapan

Chatbot bertenaga AI yang dilatih berdasarkan pola bahasa pra-pandemi mulai salah menafsirkan pertanyaan pengguna dalam konteks pasca-COVID. Pemantauan deteksi penyimpangan mengungkapkan peningkatan tajam dalam maksud yang tidak dikenali, yang mendorong pelatihan ulang model AI dengan data terbaru.

Manajemen Masalah dan Respons Insiden

Organisasi yang mengimplementasikan infrastruktur TI, aplikasi bisnis, dan sistem AI harus mengimplementasikan proses manajemen masalah terstruktur yang selaras dengan kerangka kerja manajemen layanan TI (ITSM) yang ada. Nuansa khusus AI sangat penting untuk memantau dan menanggapi insiden dalam sistem AI mereka. Elemen inti dari respons insiden khusus AI meliputi:

- ✓ Kategori triase: Mengklasifikasikan masalah berdasarkan tingkat keparahan (misalnya, gangguan layanan, pelanggaran etika, kerugian pengguna, serta masalah dengan tingkat keparahan yang lebih rendah).
- ✓ Kepemilikan dan jalur eskalasi: Menentukan tim mana yang bertanggung jawab untuk setiap kelas masalah, seperti data, model, UX, dan sebagainya, dan menguraikan proses untuk meningkatkan masalah.
- ✓ Alat pelacakan: Manfaatkan sistem seperti JIRA, ServiceNow, atau dasbor khusus untuk melacak tiket terkait model.

Insiden khusus AI, termasuk halusinasi, keluaran yang tidak terduga, atau peristiwa bias, memerlukan respons yang tepat dan disesuaikan. Tata kelola AI harus tertanam dalam rencana respons insiden yang lebih luas dan mencakup:

- ➢ Kategori insiden yang telah ditentukan sebelumnya (misalnya, etika, hukum, dan keamanan)
- ➢ Prosedur dan panduan respons yang telah ditulis sebelumnya
- ➢ Pemberitahuan segera kepada perusahaan asuransi siber untuk peristiwa yang ditanggung
- ➢ Pemberitahuan pemangku kepentingan dan pohon eskalasi
- ➢ Prosedur postmortem yang mencakup pembaruan pelatihan, dokumentasi, atau kebijakan

Respons insiden AI harus diintegrasikan ke dalam kerangka kerja respons insiden keamanan siber dan privasi organisasi secara keseluruhan, memastikan pendekatan respons insiden keseluruhan yang konsisten dan mudah dipahami.

Anatomi Insiden AI

Berikut adalah langkah-langkah yang dilakukan dalam skenario insiden AI tipikal:

- ❖ Pemicu: Pengguna/pelanggan mengajukan keluhan tentang perlakuan tidak adil.
- ❖ Analisis: Tim meninjau perilaku model dan jalur pengambilan keputusan.
- ❖ Dampak: Sistem ditemukan salah mengklasifikasikan berdasarkan dialek.
- ❖ Komunikasi: Manajemen dan kepemimpinan pada tingkat yang tepat diinformasikan dan diperbarui secara berkala.
- ❖ Tindakan: Perbaikan cepat (hotfix) diterapkan; pelatihan ulang jangka panjang dijadwalkan; pengguna/pelanggan diberitahu; pencatatan respons insiden dilakukan.
- ❖ Tata Kelola: Kartu model diperbarui, insiden dicatat untuk audit, AAR pasca-insiden dilakukan.

Pelatihan dan Pemberdayaan Pengguna

Untuk banyak sistem AI, pengguna biasanya adalah karyawan, seperti analis, agen layanan pelanggan, dan pemroses klaim, yang berinteraksi dengan model dalam alur kerja yang dibantu alat. Keberhasilan sistem AI baru membutuhkan:

- **Pelatihan berbasis peran:** Instruksi yang disesuaikan untuk berbagai kelompok pengguna (misalnya, agen lini depan vs. supervisor).
- **Pelatihan administratif:** TI dan pihak lain yang tanggung jawabnya termasuk mengelola sistem AI dan infrastruktur yang mendasarinya.
- **Penentuan ambang batas kepercayaan:** Mengajari pengguna kapan harus mempercayai sistem dan kapan harus melakukan intervensi.
- **Dokumentasi dan alat bantu kerja:** Membuat lembar contekan, alur kerja, dan panduan eskalasi dapat membantu pengguna menggunakan sistem AI sejak awal kurva pembelajaran mereka.
- **Pemberdayaan berkelanjutan:** Memperbarui pelatihan dan alat bantu kerja seiring perkembangan model dan perubahan alur kerja.

Misalnya, perusahaan pengiriman dan logistik menerapkan model optimasi rute pengiriman. Petugas pengiriman dilatih untuk meninjau saran AI dan menerapkan penggantian jika terjadi gangguan lokal (misalnya, penutupan jalan atau sakitnya pengemudi).

Untuk sistem AI yang berinteraksi langsung dengan pelanggan, implementasi harus mencakup transparansi dan panduan bagi pengguna eksternal. Hal ini mungkin meliputi:

- Pengungkapan keberadaan kemampuan AI
- Deskripsi cara kerja AI dalam bahasa yang mudah dipahami
- Mekanisme pemberitahuan dan persetujuan, termasuk tautan ke kebijakan privasi atau pemberitahuan praktik privasi (NOPP)
- Tutorial, FAQ, dan integrasi dukungan pelanggan

Keselarasan Tata Kelola Lintas Fungsi

Tata kelola AI bukanlah sesuatu yang berdiri sendiri, tetapi bergantung pada kolaborasi lintas fungsi dengan badan tata kelola lainnya. Peserta pendukung yang umum ditunjukkan pada Tabel 10.3.

Mekanisme koordinasi yang digunakan di antara badan-badan tata kelola dapat mencakup dewan AI yang bertanggung jawab tetap, ruang perang penerapan, keanggotaan lintas tata kelola, dan dasbor bersama.

Tata kelola dan manajemen sistem AI tidak boleh terisolasi; sebaliknya, keduanya harus diintegrasikan ke dalam kerangka kerja tata kelola, risiko, dan kepatuhan (GRC) yang menyeluruh. Ini berarti bahwa risiko AI harus dimasukkan dalam daftar risiko perusahaan, kontrol AI harus dimasukkan dalam audit, dan gerbang tinjauan penerapan harus diselaraskan dengan alur kerja siklus hidup tata kelola produk.

Sebagai contoh, sistem penilaian kredit berbasis AI suatu bank ditinjau sebagai bagian dari audit kepatuhan SOX triwulanan. Sistem tersebut harus memenuhi kriteria tata kelola model yang ditetapkan oleh komite risiko tingkat dewan.

Tabel 10.3	
Tata Kelola Lintas Fungsi untuk Penerapan Sistem AI	
Peran	**Tanggung jawab**
Ilmu data	Penerapan model, metrik pemantauan
TI/Pengembangan	Infrastruktur, CI/CD, skalabilitas
Legal	Kepatuhan terhadap hukum dan kontrak
Keamanan	Deteksi ancaman dan respons insiden
Tata Kelola	Keselarasan dengan nilai-nilai dan kebijakan organisasi.
Unit bisnis	Realisasi nilai, adopsi pengguna

10.3 PEMANTAUAN BERKELANJUTAN

Penyebaran bukanlah akhir dari siklus hidup sistem AI, tetapi awal dari fase operasional baru yang memperkenalkan aktivitas baru. Pemantauan berkelanjutan memastikan bahwa sistem AI tetap akurat, adil, berkinerja tinggi, dan aman.

Pemantauan juga menegaskan keselarasan dengan tujuan sistem AI, bahkan ketika sistem tersebut menghadapi data baru, lingkungan yang berubah, dan harapan yang berkembang. Bagian ini membahas strategi pemantauan, kerangka kerja tata kelola, jadwal pelatihan ulang, dan rencana pemeliharaan yang diperlukan untuk memastikan efektivitas dan kepercayaan AI yang berkelanjutan.

Dasar-Dasar Pemantauan Sistem AI

Sistem AI bersifat probabilistik dan bergantung pada data, rentan terhadap degradasi seiring waktu karena perubahan data dunia nyata. Tanpa pemantauan berkelanjutan, prediksi model AI dapat menjadi usang, bias, atau bahkan berbahaya. Pemantauan berkelanjutan memastikan bahwa:

- Kinerja tetap berada dalam ambang batas yang ditentukan
- Penyimpangan terdeteksi sejak dini
- Risiko baru dan kasus ekstrem diidentifikasi dan ditangani
- Kewajiban kepatuhan yang muncul dipenuhi

Pemantauan sistem AI merupakan kebutuhan teknis dan keharusan tata kelola. Area pemantauan umum meliputi:

- **Kualitas prediksi:** Akurasi, presisi, recall, dan kepercayaan
- **Input data:** Volume, format, anomali, nilai yang hilang, dan input berbahaya seperti injeksi prompt dan serangan peracunan model
- **Latensi dan kinerja:** Waktu respons di bawah beban yang bervariasi
- **Konsumsi sumber daya:** Konsumsi CPU/GPU, memori, penyimpanan, dan jaringan
- **Keadilan:** Perbedaan antar subkelompok dari waktu ke waktu
- **Penyimpangan:** Pergeseran dalam distribusi data atau output model
- **Metrik penggunaan:** Tingkat adopsi, pengabaian, dan perilaku pengguna

Misalnya, asisten perekrutan berbasis AI suatu organisasi memantau dan melacak tingkat penolakan berdasarkan jenis kelamin dan ras untuk mendeteksi dan mengurangi dampak yang tidak setara dari waktu ke waktu. Anomali dilaporkan kepada manajemen.

Lihat Bab 6 untuk informasi mengenai tanggung jawab yang dinyatakan dari pengembang, penyebar, dan operator, sebagaimana diuraikan dalam Undang-Undang AI Uni Eropa.

Kerangka Kerja Pemantauan Teknis

Infrastruktur dan alat pemantauan modern biasanya mencakup:

- ✓ **Pencatatan transaksi:** Merekam input, output, metadata, dan kesalahan
- ✓ **Dasbor:** Menyediakan KPI dan visualisasi secara real-time dan historis
- ✓ **Peringatan:** Dipicu ketika KPI melebihi ambang batas yang telah ditentukan
- ✓ **Log audit:** Merekam peristiwa siklus hidup sistem dan model untuk analisis forensik, peringatan insiden, dan audit kepatuhan

Alat pemantauan standar meliputi Prometheus (metrik), Grafana (dasbor), ELK stack (log), dan alat khusus AI/ML seperti Evidently, WhyLabs, atau AWS SageMaker Model Monitor.

Observabilitas sistem AI berarti suatu organisasi dapat memahami mengapa model AI menghasilkan output tertentu atau gagal dalam kondisi tertentu. Observabilitas bertujuan untuk menyediakan keterlacakan jalur prediksi, pencatatan versi dan konfigurasi model, dan korelasi output dengan input data. Observabilitas sangat penting untuk industri yang diatur dan untuk sistem yang membuat keputusan berdampak tinggi.

Mendeteksi Pergeseran dan Penurunan Kinerja Model

Seperti yang dibahas di seluruh buku ini, pergeseran model terjadi ketika data yang dilihat model dalam produksi berbeda secara signifikan dari data yang digunakan untuk melatihnya. Jenis-jenis pergeseran model meliputi:

- ➢ **Pergeseran kovariat:** Perubahan distribusi input (misalnya, perilaku pelanggan baru)
- ➢ **Pergeseran label:** Perubahan distribusi hasil (misalnya, peningkatan gagal bayar pinjaman)
- ➢ **Pergeseran konsep:** Perubahan hubungan antara input dan output (misalnya, taktik penipuan berkembang)

Teknik deteksi pergeseran meliputi:

- ❖ **Uji statistik:** *Kolmogorov-Smirnov* (K-S) dan Indeks Stabilitas Populasi (PSI)
- ❖ **Pemantauan prediksi:** Penurunan tajam dalam kepercayaan atau konsistensi
- ❖ **Model bayangan:** Model paralel yang dilatih pada data terbaru untuk perbandingan dengan model dalam produksi

Alat deteksi harus mendukung lingkungan batch dan streaming.

Misalnya, model deteksi penipuan keuangan perusahaan kartu kredit menjadi kurang efektif karena penipu beradaptasi dengan pola keputusannya—suatu bentuk pergeseran konsep.

Pemantauan Keadilan dan Etika

Setelah diterapkan, model AI dapat menunjukkan bias meskipun telah melewati pemeriksaan keadilan selama pengembangan dan pada saat rilis. Pemantauan keadilan berkelanjutan harus mencakup:

- Melacak pola prediksi di seluruh atribut sensitif dan mengamati tren yang dapat mengindikasikan bias
- Mengamati degradasi subkelompok dari waktu ke waktu

- Transparansi model AI
- Meninjau keluhan pengguna dan hasil banding—baik apa yang mereka keluhkan maupun jumlah keluhan

Pemantauan etika sistem AI dapat lebih sulit untuk diotomatisasi, tetapi pemantauan dan peringatan dapat mencakup:

- Lonjakan jumlah pengguna yang memilih keluar atau keluhan
- Peningkatan konten yang ditandai (untuk sistem pembuatan konten)
- Log eskalasi yang melibatkan penggantian atau pengembalian model

Tata kelola AI harus menerima ringkasan reguler dan memiliki wewenang untuk menyatakan insiden, yang dapat mengakibatkan penangguhan atau pelatihan ulang sistem yang bermasalah. Misalnya, model AI untuk pelengkapan otomatis email ditemukan secara konsisten menyarankan bahasa yang bias gender: "perawat" dengan "dia" dan "insinyur" dengan "dia." Audit berkelanjutan mengungkapkan pola tersebut dan mendorong pelatihan ulang dengan data yang seimbang.

Menentukan Jadwal Pemeliharaan

Meskipun pemantauan sistem AI sangat penting, hal itu saja tidak cukup. Pemeliharaan berkala juga diperlukan untuk sistem AI, termasuk:

- **Evaluasi rutin:** Penilaian kualitas model triwulanan atau bulanan
- **Pelatihan ulang terjadwal:** Membangun kembali model dengan dataset yang diperbarui; frekuensinya bergantung pada kasus penggunaan model dan volatilitas lingkungan operasi dan dapat berkisar dari mingguan hingga tahunan
- **Interval perbaikan cepat:** Patch segera untuk kegagalan mendesak
- **Tinjauan bisnis:** Tinjauan bisnis triwulanan untuk mempelajari metrik, KPI, dan insiden
- **Pembaruan ketergantungan dan kepatuhan:** Memperbarui pustaka, kerangka kerja, dan citra kontainer serta melakukan penilaian ulang kepatuhan atau peraturan secara berkala

Melatih ulang model AI berpotensi mahal, membutuhkan kurasi data, validasi, pengujian, dan penyebaran ulang. Organisasi perlu menyeimbangkan:

- Peningkatan akurasi vs. biaya infrastruktur
- Kebutuhan akan kesegaran vs. risiko ketidakstabilan
- Kewajiban hukum vs. prioritas bisnis

Alat seperti pipeline model otomatis (misalnya, MLFlow dan TensorFlow Extended—TFX) dapat membantu mengurangi biaya operasional.

Sebagai contoh, chatbot asuransi dilatih ulang setiap bulan dengan maksud yang baru diberi tag dan log umpan balik pelanggan. Lingkungan staging menguji setiap model baru sebelum penyebaran.

Penanganan Risiko yang Muncul

Karena ada begitu banyak faktor dinamis di dalam dan di sekitar sistem AI, pemantauan berkelanjutan harus mencakup indikator utama risiko sistemik, termasuk:

- ✓ Peningkatan pesat dalam kegagalan kasus ekstrem
- ✓ Penggantian prediksi oleh manusia yang berulang

- ✓ Penurunan kepercayaan pengguna atau metrik keterlibatan
- ✓ Peningkatan toksisitas output, informasi yang salah, dan halusinasi
- ✓ Peningkatan pesat dalam kasus penyalahgunaan dan penggunaan yang tidak tepat

Tim tata kelola harus mempertahankan "pemicu" yang memicu tinjauan atau penangguhan model.

Kejadian di salah satu area ini harus dianggap sebagai insiden, dengan tingkat respons insiden yang sesuai seperti yang diarahkan oleh prosedur buku panduan respons insiden. Misalnya, mesin rekomendasi produk mengalami keluhan pelanggan tentang saran yang tidak relevan atau menyinggung. Dasbor memantau menandai anomali tersebut dan membekukan pembaruan model lebih lanjut sambil menunggu pengamatan.

Pengawasan Manusia dalam Pemantauan

Karena pemantauan melibatkan banyak disiplin ilmu, tanggung jawab pemantauan harus diberikan di berbagai peran, termasuk:

- ➢ **Rekayasa AI/ML:** Kinerja, pergeseran, dan akurasi
- ➢ **Tata kelola data:** Keadilan, silsilah data, dan kepatuhan penggunaan
- ➢ **Keamanan siber:** Intelijen ancaman, kerentanan, dan upaya serangan
- ➢ **Operasi/TI:** Latensi, waktu aktif, dan pemanfaatan sumber daya
- ➢ **Etika/risiko:** Pelanggaran etika, bias, dan halusinasi
- ➢ **Audit internal:** Hasil dan tindakan yang diambil
- ➢ **Perkembangan regulasi:** Hukum baru, perubahan penegakan hukum, dan preseden hukum kasus
- ➢ **Pemeliharaan:** Operasi, log, dan pelatihan ulang

Peta kepemilikan harus mencakup pohon eskalasi dan mendukung cakupan 24/7 jika diperlukan untuk sistem kritis. Bahkan dengan otomatisasi, manusia harus tetap "terlibat" untuk memvalidasi anomali, menanggapi peringatan, dan menggunakan penilaian.

Seperti yang dinyatakan sebelumnya, pemantauan terkadang akan memaksa personel pemantauan untuk menyatakan insiden, memicu proses respons insiden. Misalnya, moderasi konten berbasis AI menandai sejumlah besar kesalahan positif untuk satire. Peninjau manusia turun tangan untuk mengkalibrasi ulang ambang batas klasifikasi dan meningkatkan pelabelan.

Pemantauan Pihak Ketiga

Sejauh ini, diskusi tentang penerapan dan pemantauan mengasumsikan model AI dioperasikan oleh organisasi yang menerapkannya. Namun, karena banyak organisasi menggunakan sistem AI yang dioperasikan oleh pihak ketiga (seringkali dalam lingkungan SaaS), beberapa aspek pemantauan mungkin tidak tersedia bagi organisasi. Sebaliknya, hanya vendor yang akan memiliki visibilitas ke aspek-aspek tertentu dari sistem AI, khususnya infrastruktur dasarnya, yang seringkali tersembunyi dalam lingkungan SaaS.

Pengurangan visibilitas dan kontrol terhadap sistem AI dan lingkungannya dapat sebagian diimbangi dengan beberapa cara, termasuk:

- ❖ Syarat dan ketentuan khusus dalam perjanjian hukum antara organisasi pelanggan dan penyedia sistem AI

- ❖ Audit berkala terhadap lingkungan operasional vendor, termasuk pengembangan, penerapan, pemantauan, respons insiden, dan operasi keamanan siber

Diasumsikan bahwa organisasi memiliki program manajemen risiko pihak ketiga, dan program tersebut tidak dijelaskan secara rinci di sini. Untuk tinjauan komprehensif tentang praktik manajemen risiko pihak ketiga, lihat buku-buku yang diterbitkan penulis tentang sertifikasi CISSP, CISA, CISM, atau CRISC.

10.4 RINGKASAN

Penerapan sistem AI dimulai dengan kesiapan pra-rilis, yang mencakup validasi bahwa model tersebut terdokumentasi dengan baik, sesuai, dan layak secara operasional. Tim pengembangan memperbarui kartu model, melakukan penilaian kesesuaian, dan menyelesaikan dokumentasi yang menjelaskan kasus penggunaan, keterbatasan, aliran data, dan pemantauan. Kesiapan produksi juga memerlukan keputusan arsitektur mengenai lingkungan penerapan, baik berbasis cloud, on-premises, atau di edge. Organisasi kemudian mempertimbangkan strategi adaptasi model seperti penyempurnaan atau generasi augmentasi pengambilan (RAG). Proses rilis juga mempersiapkan peran pengawasan manusia dan mekanisme pengembalian jika hasil yang tidak terduga muncul setelah peluncuran.

Setelah sistem secara resmi disetujui untuk produksi, praktik penerapan harus selaras dengan kebijakan organisasi, prosedur manajemen layanan TI, kerangka kerja manajemen risiko, dan standar etika. Fase ini tunduk pada protokol penerapan, gerbang persetujuan, rencana pengembalian, dan prosedur kontrol versi. Tata kelola data tetap penting, terutama dalam mengelola input data waktu nyata, menegakkan kebijakan retensi dan akses, dan memastikan kepatuhan terhadap peraturan yang berlaku. Manajemen risiko penerapan mencakup kegagalan teknis dan paparan reputasi dan etika.

Organisasi harus siap untuk mendeteksi dan menanggapi ancaman yang muncul, termasuk eksploitasi musuh, penyalahgunaan model, pelanggaran keadilan, dan intelijen ancaman. Pelatihan pengguna memastikan bahwa pengguna memahami cara berinteraksi dengan dan menafsirkan output AI dengan tepat.

Pemantauan berkelanjutan memastikan integritas sistem AI dari waktu ke waktu, memungkinkan organisasi untuk mendeteksi penyimpangan data atau kinerja, lonjakan latensi, degradasi keadilan, halusinasi, dan input serta perilaku pengguna yang tidak terduga.

Sistem pemantauan melacak distribusi input, kepercayaan model, dampak subkelompok, dan kesehatan sistem secara real-time. Kerangka kerja tata kelola mendefinisikan siapa yang bertanggung jawab untuk meninjau dan memantau data, ambang batas mana yang memicu peringatan dan insiden, dan bagaimana keputusan pelatihan ulang atau penangguhan dibuat.

Jadwal pemeliharaan mencakup siklus evaluasi reguler, protokol perbaikan cepat, dan analisis biaya-manfaat untuk pelatihan ulang. Pengawasan manusia tetap penting, baik sebagai mekanisme keselamatan real-time maupun fungsi kontrol dan tata kelola strategis.

Log pemantauan dan laporan transparansi memberikan kemampuan audit dan akuntabilitas yang diperlukan untuk penggunaan AI yang bertanggung jawab dan sesuai.

Artefak ini menangkap proses tata kelola dan aktivitas sistem AI, termasuk peristiwa penting, keputusan model, dan perilaku sistem, memungkinkan ketertelusuran bagi pemangku kepentingan internal dan regulator. Pada akhirnya, pemantauan bukan hanya tentang kontinuitas teknis tetapi juga menjaga kepercayaan publik, meminimalkan kerugian, dan memastikan bahwa sistem AI yang diterapkan tetap selaras dengan tujuan awalnya.

Tata kelola AI meluas ke lingkungan produksi dengan ketelitian yang sama seperti yang diterapkan selama pengembangan. Dengan menanamkan pengawasan, kemampuan beradaptasi, dan transparansi ke dalam setiap tahap penerapan dan pemantauan, organisasi dapat mengelola seluruh siklus operasional sistem AI dengan integritas dan kepercayaan diri.

BAB 11
MANAJEMEN DAN PENJAMINAN RISIKO AI

11.1 MANAJEMEN DAN PERAMALAN RISIKO AI

Meskipun sistem AI memperkenalkan kemampuan baru yang ampuh, sistem ini menimbulkan risiko baru yang tidak mudah ditangkap oleh pendekatan manajemen risiko perusahaan (ERM) tradisional. Sementara kerangka kerja risiko TI konvensional berfokus pada ketersediaan infrastruktur, kehilangan data, atau kegagalan sistem, risiko AI muncul dari perilaku probabilistik, keluaran yang berkembang, ketergantungan pada model pihak ketiga, dan dampak tingkat masyarakat.

Para profesional keamanan siber akan mengatakan bahwa risiko terkait AI ada pada atau di atas Lapisan 7 (dari model interkoneksi OSI), yang berarti risiko ini terkait dengan penggunaan data sosio-teknis, bukan logika sistem informasi dan infrastruktur pendukungnya.

Mengidentifikasi dan Mengkategorikan Risiko AI Secara Proaktif

Manajemen risiko AI yang efektif dimulai dengan identifikasi risiko yang jelas. Tidak seperti sistem TI statis, model AI berevolusi berdasarkan data pelatihan, umpan balik, masukan lingkungan, dan jadwal pelatihan ulang, yang berarti risiko baru dapat muncul bahkan setelah penerapan.

Dalam hal pemikiran manajemen risiko saat ini, ini tidak jauh berbeda dari manajemen risiko aplikasi dan sistem informasi tradisional. Sistem yang aman (bebas dari kerentanan yang dapat dieksploitasi) saat ini kemungkinan besar tidak akan tetap aman besok, karena para peneliti akan terus menemukan kerentanan baru. Dengan demikian, klaim bahwa suatu sistem bebas risiko paling banter hanya bersifat sementara.

Risiko AI dapat diklasifikasikan ke dalam domain berikut:

- Risiko terkait data: Bias, kualitas buruk, kebocoran, transfer data lintas batas yang dilarang, atau penggunaan data pribadi yang ilegal/tidak etis.
- Risiko terkait model: Overfitting, kurangnya transparansi, penggunaan model kotak hitam, atau kerentanan yang dapat memungkinkan serangan injeksi cepat dan serangan peracunan model.
- Risiko penerapan: Penyalahgunaan dalam konteks yang tidak dimaksudkan sejak awal, kejutan otomatisasi, dan penyimpangan kinerja.
- Risiko operasional: Konsumsi sumber daya, kegagalan integrasi, dan kegagalan berantai.
- Risiko etika dan sosial: Diskriminasi, hilangnya kendali, dan manipulasi opini publik.
- Risiko regulasi dan hukum: Ketidakpatuhan terhadap hukum perlindungan data khusus AI atau hukum perlindungan data umum yang sedang berkembang.

Ketidakpastian sistem AI, dikombinasikan dengan kecepatan dan kecanggihan yang digunakan penjahat siber untuk menyerang dan mengeksploitasi sistem informasi, seharusnya mendorong setiap organisasi untuk secara proaktif dan agresif mengidentifikasi dan

mengurangi risiko. Risiko yang melekat pada AI jauh lebih besar daripada banyak teknologi baru lainnya. Alih-alih hanya duduk dan menunggu risiko muncul, keluarlah dan carilah risiko tersebut.

Sebagai contoh, sebuah lembaga keuangan mengadopsi model AI deteksi penipuan pihak ketiga yang dilatih berdasarkan data klaim historis. Audit pasca-implementasi mengungkapkan bahwa model tersebut memiliki tingkat positif palsu yang sangat tinggi untuk pelanggan kelahiran luar negeri, yang disebabkan oleh bias dalam set pelatihan. Risiko reputasi dan hukum menyebabkan penarikan sistem dan permintaan maaf publik.

Manajemen Risiko dalam Rantai Pasokan AI

Sebagian besar organisasi tidak membangun sistem AI dari awal. Sebaliknya, mereka bergantung pada vendor, model sumber terbuka, atau layanan AI berbasis cloud. Opsi-opsi ini mungkin lebih mudah diimplementasikan, tetapi menciptakan ketergantungan pada pihak ketiga yang harus dikelola secara proaktif.

Saat pengadaan atau penerapan alat AI pihak ketiga, ketentuan kunci dalam perjanjian dapat memiliki implikasi langsung terhadap risiko. Praktik penting dalam manajemen risiko pihak ketiga (TPRM) adalah melakukan penilaian awal yang menyeluruh terhadap calon vendor dan alat-alatnya, dan kemudian menggunakan penilaian tersebut untuk menginformasikan negosiasi kontrak. Tim tata kelola harus mengidentifikasi dan mengevaluasi:

- **Ketentuan lisensi:** Komponen AI sumber terbuka (misalnya, model, kumpulan data, dan kerangka kerja) mungkin memiliki persyaratan lisensi yang ketat atau longgar (misalnya, GPL, AGPL), yang dapat menimbulkan kewajiban seperti pengungkapan kode atau batasan penggunaan.
- **Pembatasan penggunaan data:** Perjanjian vendor harus secara jelas membahas kepemilikan data, penggunaan yang diizinkan, dan penanganan data sensitif. Ini termasuk apakah vendor dapat menyimpan atau menggunakan data pelanggan untuk pelatihan ulang.
- **Transparansi dan penjelasan model:** Para profesional tata kelola harus meminta pengungkapan tentang sumber data pelatihan, silsilah model, dan metode evaluasi kinerja, serta tentang mekanisme untuk mengaudit dan menantang keputusan model.
- **Ganti rugi dan tanggung jawab:** Vendor AI sering kali menolak tanggung jawab atas segala hasil berbahaya yang dihasilkan dari keluaran model. Organisasi harus hati-hati menegosiasikan klausul tanggung jawab untuk memastikan jalan keluar jika terjadi kerusakan hukum, keuangan, atau reputasi.
- **Asuransi tanggung jawab siber:** Organisasi harus mewajibkan vendor untuk memiliki asuransi tanggung jawab siber, terutama untuk vendor yang lebih kecil, di mana pelanggaran dapat menimbulkan kesulitan keuangan yang parah.
- **Posisi keamanan dan kepatuhan:** Mintalah bukti sertifikasi atau pengesahan (misalnya, ISO/IEC 27001, ISO/IEC 42001, dan SOC 2 Tipe II) dan dokumentasi praktik pengembangan model yang aman (misalnya, sanitasi data, pengujian kerentanan, dan pengerasan model).

Sebagai contoh, model bahasa sumber terbuka yang dirilis di bawah lisensi non-komersial secara tidak sengaja diintegrasikan ke dalam platform chatbot yang berorientasi keuntungan, sehingga perusahaan tersebut berpotensi menghadapi klaim pelanggaran hak kekayaan intelektual.

Klausul Kontrak Umum yang Perlu Diperiksa

Para vendor dikenal sering menetapkan ketentuan kontrak sepihak yang memberikan sedikit hak kepada pelanggan. Berikut beberapa contoh yang harus dianggap sebagai tanda bahaya:

- Data yang dikirim ke model menjadi milik vendor.
- Data dapat disimpan untuk tujuan peningkatan model.
- Data dapat dikirim ke pihak ketiga untuk melakukan layanan.
- Output yang dihasilkan oleh model tidak dijamin akurat atau sesuai hukum.
- Pelanggan memberikan lisensi yang tidak dapat dicabut untuk menggunakan umpan balik untuk produk di masa mendatang.

Memprediksi Kerugian Sekunder dan Hilir

Komponen kunci manajemen risiko adalah antisipasi proaktif terhadap risiko masa depan dan konsekuensi yang tidak diinginkan. Konsep ini sangat relevan dengan AI, karena berbagai konsekuensi terkait AI sebagian besar masih belum dieksplorasi. Banyak konsekuensi baru terlihat ketika sistem digunakan dalam skala besar atau dalam konteks yang tidak terduga.

Penggunaan sekunder mengacu pada aplikasi keluaran AI yang berbeda dari tujuan awalnya, seringkali tanpa kesadaran atau kontrol institusional resmi. Ini dapat mencakup:

- **Penggunaan ulang input:** Menggunakan konten yang dihasilkan AI sebagai data pelatihan di tempat lain
- **Penggabungan sistem:** Menghubungkan keluaran dari satu model ke sistem lain
- **Adaptasi jahat:** Menggunakan ulang alat yang tidak berbahaya (misalnya, peningkat resolusi gambar) untuk penggunaan yang berbahaya (misalnya, deepfake)

Kerugian hilir terjadi ketika konsekuensi dari keluaran AI dirasakan oleh individu, komunitas, atau sistem yang awalnya tidak dipertimbangkan oleh pengembang atau penyebar. Kerugian ini dapat mencakup:

- ✓ **Sosial:** Misalnya, praktik perekrutan diskriminatif dari penyaringan resume algoritmik
- ✓ **Lingkungan:** Misalnya, permintaan komputasi besar-besaran untuk pelatihan AI yang menyebabkan emisi karbon berlebihan
- ✓ **Politik:** Misalnya, LLM yang secara tidak sengaja menghasilkan informasi yang salah selama pemilihan

Alat untuk memetakan kerugian hilir dapat mencakup pemetaan dampak pemangku kepentingan, pemodelan skenario rekursif, panel konsultasi eksternal, kelompok fokus, pengujian penyalahgunaan, dan red teaming.

Kerangka kerja tata kelola harus memperluas tinjauan risiko di luar pengguna dan pelanggan langsung untuk mencakup pemangku kepentingan yang terpengaruh oleh efek orde

kedua dan ketiga. Misalnya, model AI yang dilatih untuk menghasilkan esai akademis kemudian digunakan oleh siswa untuk mencontek ujian, memicu skandal plagiarisme dan pertanyaan tentang integritas institusional.

Teknik Peramalan Risiko

Praktik peramalan risiko AI melibatkan penggunaan pemodelan prediktif untuk melihat bagaimana sistem dapat berperilaku di bawah berbagai tekanan internal dan eksternal. Teknik-teknik ini meliputi:

- **Deteksi pergeseran model:** Kinerja model AI dapat menurun dari waktu ke waktu karena pergeseran data (pergeseran distribusi input) atau pergeseran konsep (perubahan variabel target). Pergeseran ini dapat meningkatkan bias, tingkat kesalahan, atau kerugian.
- **Mengantisipasi lingkaran umpan balik:** Sistem AI yang memengaruhi lingkungan tempat mereka belajar (misalnya, mesin rekomendasi dan kepolisian prediktif) dapat menciptakan lingkaran umpan balik yang saling memperkuat yang memperkuat bias atau menyebabkan perilaku yang tak terkendali.
- **Peramalan risiko melalui simulasi:** Kembaran digital, data sintetis, atau lingkungan pengguna yang disimulasikan dapat membantu organisasi mengantisipasi bagaimana sistem AI akan berperilaku dalam skenario dunia nyata dan yang bersifat antagonis.

Pertimbangkan untuk menggunakan alat deteksi pergeseran statistik seperti uji KS dan divergensi KL untuk memicu pelatihan ulang atau peringatan ketika ambang batas dilampaui. Misalnya, algoritma kepolisian prediktif suatu komunitas mengirimkan lebih banyak petugas ke lingkungan tertentu, yang mengakibatkan lebih banyak penangkapan, yang pada gilirannya memperkuat lokasi tersebut sebagai berisiko tinggi dalam data yang dilatih ulang.

Menanamkan Manajemen Risiko ke dalam Struktur Tata Kelola

Manajemen risiko AI adalah proses siklus hidup yang merupakan komponen kunci dari tata kelola AI. Manajemen risiko AI harus diintegrasikan ke dalam setiap fase siklus hidup pengembangan:

- **Inisiasi:** Tentukan tingkat risiko kasus penggunaan; Konsultasi dengan pemangku kepentingan lintas fungsi
- **Desain:** Lakukan penilaian risiko model, termasuk perkiraan kerugian; identifikasi asal usul data pelatihan
- **Pengembangan:** Gunakan algoritma yang dapat dijelaskan dan diuji; lacak silsilah
- **Pengujian:** Sertakan pengujian lawan, red teaming, dan evaluasi potensi penyalahgunaan
- **Penerapan:** Tetapkan prosedur rollback dan ambang batas insiden
- **Pasca-penerapan:** Pantau penyimpangan, log audit, umpan balik pengguna, dan laporan insiden

Sertakan risiko terkait AI dalam register risiko ERM dan gunakan proses manajemen risiko yang sama untuk memproses semua risiko, termasuk risiko AI.

Penetapan peran terkait risiko yang jelas meningkatkan tata kelola dan memungkinkan ketertelusuran. Pertimbangkan untuk menunjuk peran-peran berikut:

- **Petugas risiko AI atau kepala petugas risiko:** Bertanggung jawab atas postur risiko tingkat perusahaan
- **Pemilik bisnis/eksekutif senior:** Bertanggung jawab atas keputusan penanganan risiko
- **Pemilik sistem:** Bertanggung jawab atas model individual
- **Pemimpin audit:** Mengawasi pengujian dan ritme peninjauan
- **Petugas perlindungan data (DPO)/penasihat hukum:** Mengawasi kepatuhan terhadap perlindungan data, kekayaan intelektual, dan persyaratan peraturan khusus AI

Program tata kelola AI sering menggunakan alat bantu untuk mengatur prosesnya, termasuk:

- Kerangka Kerja Manajemen Risiko AI NIST (AI RMF)
- ISO/IEC 42001 (Standar Sistem Manajemen AI)
- Kartu model yang berfokus pada risiko
- Templat peramalan kerugian
- Deteksi risiko otomatis dalam pipeline MLOps

Seiring dengan kematangan kemampuan dan ekosistem AI, para profesional tata kelola harus memprioritaskan alat bantu yang memungkinkan deteksi risiko, transparansi, dan remediasi yang berkelanjutan. Dalam lingkungan dengan komponen AI berlapis, peramalan risiko harus menjadi lebih terperinci, berkelanjutan, dan adaptif.

Dokumentasi dan Jaminan

Dokumentasi yang lengkap dan akurat mendukung komunikasi risiko internal, audit eksternal, investigasi insiden, pemecahan masalah model, dan tinjauan peraturan. Dokumentasi dan artefak catatan harus mencakup:

- Daftar risiko AI (catatan bisnis yang dinamis dan selalu diperbarui)
- Penilaian model pihak ketiga
- Catatan asal usul data dan persetujuan
- Catatan silsilah data
- Penilaian dampak (misalnya, Penilaian Dampak Algoritma, DPIA)
- Catatan manajemen risiko, termasuk keputusan penanganan risiko, rencana perbaikan, dan justifikasi risiko residual
- Kartu model dan sistem (dokumentasi ringkas yang menjelaskan tujuan, masukan, keterbatasan, dan hasil evaluasi)

Poin-poin yang perlu dicakup untuk tinjauan jaminan model AI harus mencakup:

- Mode kegagalan yang diketahui dari sistem AI
- Bagaimana risiko dimitigasi, dan risiko residual apa yang tersisa
- Siapa yang bertanggung jawab atas setiap kelas risiko
- Seberapa sering kontrol risiko diuji
- Bagaimana kegagalan atau kerugian yang ditemukan diselidiki, dikomunikasikan, dan diperbaiki

Pertimbangan Regulasi

Hukum AI global dan regional diberlakukan lebih cepat daripada yang dapat dibahas dalam buku ini. Banyak dari hukum AI global ini memerlukan proses manajemen risiko formal. Contoh persyaratan meliputi:

- ✓ **Undang-Undang AI Uni Eropa:** Mewajibkan penyedia sistem AI berisiko tinggi untuk melakukan penilaian risiko, memelihara log, dan menyerahkan sistem untuk penilaian kesesuaian.
- ✓ **Undang-Undang Kecerdasan Buatan dan Data Kanada (AIDA):** Mewajibkan penilaian dampak dan pengungkapan publik untuk sistem AI berdampak tinggi (pada saat penulisan, undang-undang ini belum diberlakukan).
- ✓ **Perintah Eksekutif AS tentang AI (2025):** Mendorong lembaga untuk memastikan bahwa sistem AI "bebas dari bias ideologis atau agenda sosial yang direkayasa."
- ✓ **Tindakan Sementara Tiongkok untuk Pengelolaan Layanan AI Generatif:** Mewajibkan klasifikasi sistem AI dan pengawasan berdasarkan risiko dan kategori, serta pelabelan wajib konten yang dihasilkan AI.

Organisasi yang gagal memprediksi dan mengelola risiko dengan benar dapat dikenakan sanksi, perintah pengadilan, atau tanggung jawab perdata.

11.2 AUDIT DAN PENGUJIAN SISTEM AI

Untuk memastikan AI yang bertanggung jawab, tangguh, dan patuh, organisasi harus melampaui pengembangan model dan secara aktif menilai bagaimana sistem AI berkinerja di dunia nyata. Penilaian ini memerlukan proses audit, pengujian keamanan (*red teaming*), pemodelan ancaman, dan pengujian keamanan yang disengaja dan berulang. Masing-masing memberikan wawasan yang berbeda tentang kinerja, keandalan, keamanan, dan kerentanan sistem AI. Aktivitas ini membantu mengungkap perilaku yang tidak diinginkan, celah kepatuhan peraturan, kelemahan pihak lawan, dan masalah etika yang mungkin tidak terdeteksi sampai terjadi kerugian.

Audit Kinerja dan Validasi Model

Audit kinerja berfokus pada pengukuran seberapa efektif sistem AI memenuhi tujuan yang dimaksudkan di berbagai dimensi, termasuk akurasi, keadilan, ketahanan, kinerja, dan generalisasi. Audit ini memvalidasi atau membantah klaim yang dibuat selama pengembangan dan mendeteksi degradasi dari waktu ke waktu. Dapat dikatakan, pemantauan berkelanjutan seharusnya mendeteksi degradasi lebih awal daripada audit.

Audit dimulai dengan definisi kriteria kinerja yang dapat diterima yang ditetapkan pada fase desain sistem AI. Kriteria ini harus mencakup:

- ➢ **Keselarasan bisnis:** Misalnya, akurasi deteksi penipuan
- ➢ **Sensitif konteks:** Misalnya, toleransi negatif palsu dalam diagnosis medis
- ➢ **Berbasis informasi pemangku kepentingan:** Misalnya, persetujuan yang transparan dan eksplisit
- ➢ **Keadilan dan kurangnya bias:** Misalnya, keadilan di seluruh kelompok demografis
- ➢ **Presisi dan kepercayaan:** Misalnya, akurasi klasifikasi dan pengambilan keputusan yang konsisten
- ➢ **Ketahanan dan daya tahan:** Kinerja yang dapat diterima bahkan di bawah beban dan dengan kasus-kasus ekstrem

Misalnya, model penilaian kredit memenuhi ambang batas akurasi selama pengembangan. Namun, setelah penerapan, audit mengungkapkan perbedaan kinerja ketika diterapkan pada pelamar di daerah pedesaan, yang mendorong evaluasi ulang input fitur dan pelatihan ulang model.

Auditor dan pemilik sistem harus memperhitungkan kemungkinan bahwa model AI sering berperilaku berbeda di lingkungan produksi daripada di lingkungan pengembangan dan pengujian yang terkontrol. Ketika sistem produksi gagal memenuhi kriteria yang dinyatakan, auditor dapat mempertimbangkan untuk menguji model AI yang sama di lingkungan pengujian untuk menentukan apakah model tersebut berperilaku berbeda. Jika demikian, hal ini harus mendorong penyelidikan untuk memahami perbedaannya.

Red Teaming untuk Ketahanan terhadap Serangan

Red teaming adalah proses terstruktur dan serangkaian teknik untuk mensimulasikan serangan musuh dan skenario penyalahgunaan terhadap sistem target. Red teaming sangat penting untuk mengidentifikasi kerentanan yang mungkin terlewatkan oleh tim internal, serta untuk memastikan ketahanan dalam kondisi terburuk.

Ada beberapa jenis penilaian tim merah yang menargetkan sistem AI, termasuk:

- Pengujian injeksi prompt dan jailbreak untuk model bahasa besar
- Serangan penghindaran (memodifikasi input untuk menyebabkan kesalahan klasifikasi)
- Serangan peracunan (menyuntikkan data yang rusak ke dalam pipeline pelatihan atau penyempurnaan)
- Penyalahgunaan fungsi (menggunakan alat untuk tujuan yang tidak diinginkan atau tidak etis)
- Pengujian rekayasa sosial (misalnya, phishing melalui chatbot)

Tim merah juga harus mengevaluasi ekosistem AI yang lebih luas, termasuk infrastruktur pendukung, API, lapisan integrasi, dan pipeline data. Banyak serangan mengeksploitasi infrastruktur, antarmuka, dan dependensi, bukan hanya parameter model.

Anggota tim merah harus mencakup pakar domain yang memahami sistem AI dan ancaman umum, pakar kebijakan yang akan memahami penyalahgunaan, dan penguji eksternal yang membawa objektivitas. Tim merah harus menargetkan sistem AI itu sendiri dan personel yang membangun, mengelola, dan menggunakannya.

Pemodelan Ancaman untuk Permukaan Serangan Khusus AI

Teknik pemodelan ancaman tradisional harus diperluas untuk memperhitungkan permukaan serangan khusus AI, seperti inversi model, kebocoran data pelatihan, dan serangan adversarial.

Versi modifikasi dari *Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, Elevation of Privilege* (STRIDE) dapat digunakan untuk mengidentifikasi ancaman terhadap sistem AI, termasuk:

- Peracunan data (perusakan)
- Pencurian model (pengungkapan informasi)
- Rantai API yang tidak sah (peningkatan hak akses)
- Injeksi prompt (spoofing dan perusakan)

Metodologi pemodelan ancaman lainnya meliputi:

- MITRE ATLAS (*Adversarial Threat Landscape for Artificial-intelligence Systems*): Basis pengetahuan untuk ancaman adversarial terhadap AI
- Analisis Rantai Serangan: Memetakan langkah penyerang dari pengintaian hingga eksploitasi
- Pohon Serangan: Memvisualisasikan jalur menuju kompromi
- TARA (*Threat Analysis and Risk Assessment*): Diadaptasi untuk sistem AI
- PASTA (*Process for Attack Simulation and Threat Analysis*): Dapat diperluas untuk memodelkan skenario dampak bisnis untuk penyalahgunaan AI

Selain pemodelan ancaman, perburuan ancaman adalah aktivitas penting yang digunakan untuk secara proaktif mencari tanda-tanda Intrusi, penyusup, penyalahgunaan, dan indikator kompromi. Alih-alih menunggu insiden terjadi, perburuan ancaman secara proaktif mencari tanda-tanda serangan, yang seringkali mengarah pada deteksi lebih awal dan dapat membantu mengurangi kerusakan dan kerugian.

Sebagai contoh, model AI generatif yang berhadapan dengan publik terancam oleh upaya terkoordinasi untuk mengekstrak data pelatihannya. Pemodelan ancaman mengungkapkan bahwa pengambilan sampel berulang dengan perintah yang ditargetkan dapat menghasilkan fragmen data pelatihan, yang mendorong organisasi untuk menerapkan kontrol pembatasan dan penandaan tambahan.

Pemodelan Ancaman di Seluruh Siklus Hidup

Pemodelan ancaman harus dilakukan pada berbagai titik dalam pengembangan sistem AI, termasuk:

- **Fase desain:** Mengidentifikasi fitur dan komponen yang rentan
- **Fase pra-implementasi:** Menguji pertahanan yang diimplementasikan
- **Fase pasca-implementasi:** Beradaptasi dengan ancaman dan konteks yang terus berkembang

Pengujian Keamanan untuk Kerahasiaan dan Integritas

Audit sistem AI harus melampaui kinerja dan kualitas keluaran untuk mencakup seberapa efektif sistem tersebut melindungi integritas data, privasi pengguna, dan kerahasiaan sistem. Pengujian keamanan mengevaluasi bagaimana sistem AI berperilaku dalam kondisi berbahaya dan abnormal. Teknik keamanan khusus AI meliputi:

- **Pengujian penetrasi:** Mensimulasikan upaya intrusi ke dalam sistem AI, pipeline, dan API
- **Pengujian ekstraksi model:** Menilai kemudahan rekayasa balik atau penyalinan model AI
- **Serangan inferensi:** Menentukan seberapa mudah data pelatihan dapat disimpulkan dari output

- **Inferensi keanggotaan:** Menguji apakah penyerang dapat menebak apakah data spesifik digunakan untuk melatih model
- **Pengujian re-identifikasi:** Memvalidasi anonimisasi data pelatihan untuk menentukan apakah data tersebut dapat diidentifikasi ulang
- **Pengujian ketahanan terhadap serangan:** Memeriksa bagaimana model berperilaku di bawah gangguan atau manipulasi input yang dilakukan oleh penyerang

Organisasi yang menggunakan pipeline integrasi berkelanjutan/penyebaran berkelanjutan (CI/CD) harus mengamankan lingkungan ini untuk mencegah perusakan, pembaruan yang tidak sah, atau penyebaran model AI bayangan. Pengamanan CI/CD utama meliputi:

- ✓ Pemantauan proses check-out dan check-in
- ✓ Pembatasan akses ke kode sensitif berdasarkan prinsip "perlu diketahui"
- ✓ Penandatanganan kode dan kontrol versi artefak model
- ✓ Kontainerisasi dan sandboxing selama pelatihan model
- ✓ Pengembangan dan pemantauan kontrol akses berbasis kebijakan untuk pemicu pelatihan ulang
- ✓ Pencatatan audit untuk akses data dan peristiwa penyebaran model

Sebagai contoh, sebuah perusahaan e-commerce menemukan bahwa seorang karyawan nakal telah menyebarkan model AI yang tidak disetujui ke dalam pipeline. Pengujian keamanan mengungkapkan kurangnya pemeriksaan integritas dan kontrol akses yang lemah dalam rantai CI/CD.

Pengujian Keamanan Full-stack Sangat Penting

Pengujian keamanan harus mencakup seluruh tumpukan teknologi AI, termasuk keamanan fisik, jaringan, sistem operasi, sistem manajemen basis data, server aplikasi, sistem penyimpanan, layanan jaringan, dan komponen relevan lainnya.

Audit Regulasi dan Kepatuhan

Audit berkala terhadap sistem AI membantu mengkonfirmasi kepatuhan sistem tersebut terhadap hukum, perjanjian hukum, kebijakan, dan standar etika yang berlaku. Audit ini memberikan ketertelusuran, akuntabilitas, dan pembelaan jika terjadi pengawasan regulasi. Tujuan utama audit kepatuhan meliputi:

- ➢ **Memverifikasi kepatuhan terhadap kerangka kerja manajemen risiko AI:** Misalnya, NIST AI RMF dan ISO/IEC 42001
- ➢ **Mengevaluasi kontrol yang diperlukan:** Misalnya, mitigasi bias dan penilaian dampak
- ➢ **Mengkonfirmasi keberadaan dan kualitas dokumentasi:** Misalnya, kartu model, lembar data, dan kebijakan penggunaan
- ➢ **Mengkonfirmasi keberadaan dan kualitas catatan bisnis:** Misalnya, register risiko, keputusan penanganan risiko, pemeliharaan berkala, dan permintaan opt-out
- ➢ **Mengkonfirmasi penerapan hak opt-out yang tepat**
- ➢ **Mengkonfirmasi keberadaan dan efektivitas pengamanan human-in-the-loop**

Persyaratan audit terpilih berdasarkan yurisdiksi meliputi:

- **Undang-Undang AI Uni Eropa:** Mewajibkan sistem AI berisiko tinggi untuk menyertakan sistem manajemen mutu dan menjalani penilaian kesesuaian dan pemantauan pasca-pasar.
- **AIDA Kanada:** Mewajibkan penilaian dampak dan langkah-langkah transparansi (pada saat penulisan ini, undang-undang ini belum diberlakukan).
- **Badan-badan federal AS:** Didorong untuk mengadopsi kerangka kerja tata kelola AI melalui Perintah Eksekutif dan panduan OMB yang berlaku.

Langkah-langkah untuk mempersiapkan audit sistem AI meliputi:

- Mempertahankan satu sumber kebenaran untuk dokumentasi.
- Memastikan log audit tidak dapat diubah dan diberi cap waktu.
- Melakukan uji coba atau simulasi audit internal.
- Menetapkan pemilik yang bertanggung jawab untuk setiap area kepatuhan.
- Berikan pelatihan kepada pemilik usaha yang bertanggung jawab mengenai perilaku yang tepat saat berinteraksi dengan auditor.

Pengujian Waktu Nyata dan Pemantauan Berkelanjutan

Penilaian berkala tidak hanya penting tetapi juga tidak cukup. Organisasi harus menerapkan mekanisme jaminan berkelanjutan untuk mendeteksi risiko yang muncul dan penurunan kinerja secara waktu nyata. Praktik pemantauan berkelanjutan meliputi:

- Metrik latensi dan throughput untuk mendeteksi masalah kinerja
- Deteksi pergeseran dalam distribusi input dan output
- Deteksi anomali untuk perilaku langka atau tidak terduga
- Lingkaran umpan balik pengguna untuk mengungkap masalah keamanan dan etika
- Peringatan insiden untuk mengetahui insiden saat terjadi
- Ambang batas peringatan insiden untuk pemicu remediasi otomatis
- Pemindaian keamanan untuk secara proaktif mengidentifikasi kerentanan yang baru ditemukan
- Perburuan ancaman untuk secara proaktif mendeteksi tanda-tanda penyalahgunaan dan intrusi

Misalnya, sistem AI perawatan kesehatan yang memantau tanda-tanda vital mulai menandai banyak positif palsu yang tinggi. Pemantauan berkelanjutan mengungkapkan pergeseran input bertahap yang dihasilkan dari pembaruan firmware pada perangkat medis, yang mendorong organisasi untuk memulai siklus pelatihan ulang.

Seperti yang dinyatakan di tempat lain dalam bab ini, pengujian dan pemantauan seluruh tumpukan teknologi tidak hanya penting tetapi juga dianggap sebagai praktik penting untuk sistem informasi apa pun.

Mengintegrasikan Pengujian ke dalam Siklus Hidup AI

Audit dan pengujian tidak hanya dilakukan pada lingkungan sistem AI produksi, tetapi harus disematkan ke dalam setiap tahap desain, pengembangan, dan alur kerja penerapan sistem AI. Audit tidak hanya menguji sistem AI itu sendiri, tetapi juga kebijakan, proses bisnis, dan alat yang digunakan untuk mendesain, membangun, menguji, menerapkan, dan memantaunya. Poin-poin penting untuk audit siklus hidup AI meliputi:

- **Fase desain:** Tinjauan dan persetujuan desain, penilaian risiko, pemodelan ancaman, dan perencanaan skenario red teaming
- **Fase pengembangan:** Pengujian unit untuk kode model, rangkaian pengujian adversarial, kontrol akses repositori, dan permintaan serta persetujuan perubahan fitur
- **Fase pra-deployment:** Audit kinerja ujung-ke-ujung dan tinjauan peraturan
- **Fase deployment:** Pengujian canary dan pengamanan rollback
- **Fase pasca-deployment:** Pemantauan berkelanjutan, perencanaan respons insiden, dan contoh respons insiden
- **Fase penghentian:** Pengarsipan data, penonaktifan model, pembuangan, dan migrasi ke sistem lain (penghentian sistem AI dibahas dalam Bab 12)

Beberapa organisasi menerapkan model gerbang untuk memastikan bahwa pengujian selesai sebelum proyek dapat dilanjutkan. Tahapan-tahapan tersebut mungkin meliputi:

- **Tahapan 1:** Penilaian risiko dan tinjauan desain selesai, menyetujui dimulainya pengembangan
- **Tahapan 2:** Pengujian lawan dan audit bias lulus, menyetujui promosi ke penggunaan produksi
- **Tahapan C:** Pemantauan operasional dan rencana pengembalian (rollback) divalidasi

Peran dan Tanggung Jawab Audit AI

Artefak tata kelola AI, atau piagam program audit yang terdokumentasi, harus secara jelas mendefinisikan kepemilikan dan akuntabilitas untuk kegiatan audit dan pengujian. Peran umum meliputi:

- ✓ **Petugas risiko AI:** Mengawasi strategi jaminan di seluruh sistem
- ✓ **Pemilik model:** Bertanggung jawab atas hasil pengujian spesifik sistem
- ✓ **Pemimpin audit internal:** Mengkoordinasikan kegiatan audit internal dan pihak ketiga
- ✓ **Koordinator tim merah:** Merancang dan melaksanakan evaluasi lawan
- ✓ **Insinyur keamanan:** Melakukan pengujian penetrasi, integritas, dan pemodelan ancaman
- ✓ **Penasihat etika:** Memandu evaluasi dampak sosial

Penugasan peran-peran ini memastikan upaya pengujian tidak hanya terisolasi dalam tim teknis saja, tetapi juga mencerminkan tata kelola lintas fungsi.

Dokumentasi Audit dan Penyimpanan Bukti

Piagam audit, rencana, hasil, laporan pengujian, dan data pemantauan harus disimpan dengan cara yang mendukung akuntabilitas dan penyelidikan di masa mendatang. Jadwal penyimpanan catatan organisasi harus menentukan periode penyimpanan untuk setiap jenis catatan bisnis, setelah itu harus dimusnahkan sesuai dengan kebijakan organisasi. Artefak dokumentasi audit meliputi:

- ➢ Piagam audit yang mendefinisikan program audit dan peran serta tanggung jawab audit
- ➢ Laporan audit dengan temuan, tingkat keparahan, dan rekomendasi
- ➢ Log pengujian tim merah dan skenario lawan yang diuji
- ➢ Model ancaman dengan jalur serangan dan mitigasi
- ➢ Catatan pemodelan ancaman

- Protokol pengujian dan set pengujian yang dapat direproduksi
- Riwayat versi model dan metadata penyebaran
- Bukti tindakan perbaikan yang diambil

Penyimpanan harus memenuhi kewajiban peraturan, hukum, dan etika, termasuk kontrol akses, jadwal penyimpanan, enkripsi, dan minimalisasi data. Organisasi harus berkonsultasi dengan penasihat hukum mereka untuk menentukan penerapan persyaratan penyimpanan.

Alat dan Otomatisasi Audit dan Pengujian

Pengujian manual tidak efektif untuk portofolio sistem AI yang terus berkembang, dan juga tidak efektif untuk aplikasi bisnis tradisional non-AI. Organisasi harus berinvestasi dalam alat untuk mengotomatisasi bagian-bagian penting dari alur kerja audit dan pengujian. Contoh alat pengujian khusus AI meliputi:

- Kerangka kerja audit bias: Misalnya, Aequitas dan Fairlearn
- Platform red teaming: Misalnya, alat Azure AI Red Team dari Microsoft
- Pemindai keamanan: Untuk menguji lingkungan pengembangan, pengujian, dan produksi AI
- Generator kartu model dan lembar data
- Dasbor pemantauan: Misalnya, Evidently AI, Arize, dan WhyLabs

Pemilihan alat harus selaras dengan kompleksitas sistem, tingkat risiko, dan kemampuan internal.

11.3 MANAJEMEN INSIDEN DAN PENGUNGKAPAN

Di dunia keamanan siber, secara umum dipahami bahwa, terlepas dari upaya terbaik untuk mencegah peristiwa dan insiden, hal tersebut tetap akan terjadi—membutuhkan kemampuan deteksi dan respons insiden yang efektif. Demikian pula, terlepas dari kontrol risiko dan pengujian yang ketat, sistem AI mungkin masih menghasilkan keluaran yang berbahaya, menyesatkan, atau tidak diinginkan setelah penerapan, sehingga manajemen insiden menjadi komponen penting dari tata kelola AI. Pendekatan yang kuat tidak hanya membatasi kerugian ketika insiden terjadi tetapi juga mendukung pembelajaran organisasi, kewajiban transparansi, dan kepatuhan terhadap peraturan yang ada dan yang sedang berkembang.

Mengidentifikasi dan Mengelola Insiden AI

Insiden AI mengacu pada perilaku, hasil, atau interaksi yang tidak terduga yang melibatkan sistem AI yang menyebabkan, atau berpotensi menyebabkan, kerugian, gangguan operasional, atau masalah regulasi.

Contoh insiden AI meliputi:

- Sistem rekomendasi yang menampilkan konten berbahaya atau ilegal
- Chatbot yang memberikan saran medis yang salah
- Model klasifikasi yang salah memberi label pada atribut demografis yang sensitif
- Sistem otonom yang gagal beroperasi dengan aman dalam kondisi kasus ekstrem
- Model generatif yang menghasilkan keluaran yang melanggar IP, beracun, atau tidak akurat secara faktual

- Penurunan kinerja yang tidak terduga karena pergeseran konsep atau perubahan lingkungan

Respons insiden AI harus dimodelkan berdasarkan proses manajemen insiden TI dan keamanan siber yang sudah mapan dan diadaptasi untuk sifat probabilistik dan buram dari sistem AI. Organisasi harus mempertimbangkan untuk memberlakukan kerangka kerja respons insiden tunggal yang mencakup insiden terkait TI, keamanan siber, privasi, dan AI.

Langkah-langkah utama dalam respons insiden adalah sebagai berikut:

1. **Deteksi:** Didukung oleh berbagai aktivitas pemantauan, deteksi mengidentifikasi perilaku anomali atau berbahaya melalui peringatan, pemantauan, atau laporan pengguna
2. **Triase:** Penilaian tingkat keparahan, cakupan, dan potensi dampak insiden
3. **Komunikasi:** Memberi informasi kepada pemangku kepentingan internal dan eksternal
4. **Investigasi:** Menentukan akar penyebab, termasuk model dan silsilah data
5. **Pengendalian:** Mengisolasi atau menonaktifkan sistem atau komponen yang terpengaruh, mencegah penyebarannya atau keberlanjutannya
6. **Pemberantasan:** Menerapkan tindakan korektif, seperti pelatihan ulang model, pembaruan data, atau perubahan kebijakan
7. **Pemulihan:** Mengembalikan sistem ke keadaan sebelum insiden (tetapi dengan perubahan yang diberlakukan untuk mencegah insiden serupa)
8. **Tinjauan pasca-tindakan (AAR):** Melakukan analisis akar penyebab untuk lebih memahami penyebab sebenarnya dari insiden tersebut, memungkinkan perbaikan jangka panjang yang efektif
9. **Dokumentasi:** Memperbarui kebijakan, log, dan jejak audit

Sebagai contoh, asisten dukungan pelanggan bertenaga AI generatif mulai memberikan instruksi pengembalian yang salah kepada pelanggan karena sebuah prompt Terjadi kebocoran data dari lingkungan pengujian. Triage mengidentifikasi sumbernya sebagai aliran input yang salah arah. Model tersebut untuk sementara dihentikan sementara, dilatih ulang, dan divalidasi ulang dengan filter input yang diperbarui.

Tidak Semua Insiden Merupakan Kegagalan

Beberapa insiden bukan berasal dari bug, tetapi dari tujuan yang tidak selaras. Misalnya, model yang dilatih untuk memaksimalkan keterlibatan mungkin secara tidak sengaja belajar untuk mempromosikan konten yang sensasional atau polarisasi. Oleh karena itu, respons insiden harus mempertimbangkan niat versus hasil.

Mengkategorikan dan Meningkatkan Insiden

Insiden sangat bervariasi dalam tingkat keparahan dan dampaknya; oleh karena itu, tidak semua insiden memerlukan tingkat respons atau pengungkapan yang sama. Kategorisasi yang jelas memfasilitasi upaya triase, menentukan komunikasi dan peningkatan, serta memprioritaskan perbaikan. Skala keparahan tipikal mencakup tingkat-tingkat berikut:

- **Tingkat 1—Kritis:** Kerugian yang meluas, pelanggaran peraturan, atau kegagalan sistemik
- **Tingkat 2—Mayor:** Dampak signifikan pada pengguna atau kerusakan model dalam fungsi inti
- **Tingkat C—Sedang:** Masalah terisolasi dengan paparan terbatas
- **Tingkat 4—Minor:** Tidak ada dampak eksternal; ditangani oleh pemantauan internal

Opsi peningkatan untuk insiden dengan tingkat keparahan lebih tinggi meliputi:

- Pemberitahuan eksekutif dan pembaruan berkala
- Pemberitahuan regulator eksternal
- Penangguhan hukum atas artefak terkait
- Perlindungan catatan melalui hak istimewa pengacara-klien
- Penangguhan atau pengembalian model

Organisasi yang menggunakan alur kerja otomatis dapat mengkonfigurasinya untuk memicu aturan peningkatan berdasarkan metadata insiden.

Mendokumentasikan Insiden, Masalah, dan Risiko

Meskipun diwajibkan oleh beberapa peraturan, mendokumentasikan semua peristiwa, insiden, masalah, dan risiko yang terkait dengan sistem AI dan sistem non-AI merupakan praktik penting. Dokumentasi yang efektif memastikan ketertelusuran, kesiapan regulasi, dan peningkatan memori organisasi. Pencatatan yang baik juga memungkinkan analisis tren dan mitigasi proaktif terhadap pola yang berulang.

Praktik penting untuk mendokumentasikan insiden meliputi:

- Tanggal/waktu deteksi
- Sumber deteksi (sistem pemantauan, laporan pengguna, audit)
- Sistem, layanan, atau model yang terpengaruh
- Deskripsi perilaku dan konsekuensi yang diamati
- Klasifikasi tingkat keparahan
- Tindakan respons langsung yang diambil (dan oleh siapa)
- Akar penyebab dan faktor yang berkontribusi
- Langkah-langkah mitigasi dan verifikasi
- Status dan tanggal penyelesaian

Penggunaan taksonomi standar (misalnya, jenis insiden, komponen sistem, dan jenis kerugian) membantu memfasilitasi analisis, pemahaman lintas tim, metrik, dan pelaporan.

Biasanya organisasi mencatat satu atau lebih risiko dalam register risiko ketika suatu insiden mengungkapkan risiko yang sebelumnya tidak dipertimbangkan. Hal ini juga berlaku sebaliknya: setiap insiden harus ditelusuri ke risiko yang relevan dalam register risiko AI organisasi yang sudah ada pada saat insiden terjadi.

Risiko Sisa Setelah Perbaikan

Risiko jarang dihilangkan sepenuhnya. Saat mendokumentasikan suatu insiden, catat apakah perbaikan yang diterapkan mengurangi atau menghilangkan risiko, dan apakah risiko sisa tetap berada dalam ambang batas yang dapat diterima.

Berkolaborasi dengan Pemangku Kepentingan untuk Memahami Akar Penyebab

Memahami alasan mengapa insiden AI terjadi umumnya membutuhkan pendekatan multidisiplin. Akar penyebab teknis mungkin terkait dengan keputusan desain, pilihan kebijakan, dan kesalahpahaman kontekstual.

Akar penyebab umum insiden terkait AI meliputi:

- ✓ **Kerapuhan model:** Ketidakmampuan untuk menangani kasus-kasus ekstrem atau input yang tidak terduga
- ✓ **Kurangnya ketahanan:** Gangguan kecil menyebabkan output yang tidak dapat diprediksi
- ✓ **Data pelatihan yang tidak mencukupi atau berkualitas buruk**
- ✓ **Tujuan yang tidak selaras:** Misalnya, mengoptimalkan keterlibatan vs. keselamatan pengguna
- ✓ **Ketergantungan berlebihan pada otomatisasi atau kurangnya pengawasan manusia:** Pengawasan manusia yang tidak memadai atau kurangnya keterlibatan manusia
- ✓ **Pengujian yang tidak memadai, terutama dalam kondisi yang merugikan atau kondisi dunia nyata**
- ✓ **Penyimpangan model atau data**
- ✓ **Kegagalan dalam desain yang cepat untuk AI generatif**

Tinjauan pasca-insiden harus diformalkan dan mencakup banyak, jika tidak semua, pemangku kepentingan sistem AI. Peran yang disarankan adalah sebagai berikut:

- ➢ **Ilmuwan data:** Menjelaskan perilaku model dan silsilah data pelatihan
- ➢ **Insinyur perangkat lunak:** Menyelidiki log sistem dan lapisan integrasi
- ➢ **Keamanan siber:** Memahami dan menjelaskan kerentanan dan ancaman terkait
- ➢ **Privasi:** Mengevaluasi kegagalan hak subjek data
- ➢ **Penasihat etika dan hukum:** Mengevaluasi dampak peraturan dan masyarakat
- ➢ **Staf operasional:** Mengidentifikasi dampak bisnis atau pelanggan hilir
- ➢ **Pengguna akhir (jika sesuai):** Memberikan pemahaman kontekstual tentang kerugian
- ➢ **Pakar eksternal:** Membawa objektivitas pada evaluasi

Misalnya, model pemeliharaan peralatan prediktif gagal menandai kegagalan komponen utama tepat waktu. Tinjauan lintas fungsi mengungkapkan bahwa model tersebut dilatih secara eksklusif pada kondisi operasi normal, sehingga mengakibatkan adanya titik buta terhadap mode kegagalan yang jarang terjadi namun kritis.

Menghindari Budaya Saling Menyalahkan

Tujuan analisis insiden adalah untuk belajar dan meningkatkan diri, bukan untuk mengidentifikasi dan menghukum yang bersalah. Organisasi yang mendorong keamanan

psikologis dalam tinjauan pasca-insiden akan mendorong berbagi secara terbuka dan peningkatan proses.

Transparansi dan Pengungkapan Publik

Semakin sering, insiden keamanan dan privasi tidak dapat lagi disembunyikan. Regulasi di banyak yurisdiksi mewajibkan pengungkapan insiden privasi, dan semakin sering, insiden AI. Transparansi adalah prinsip dasar tata kelola AI, dan merupakan persyaratan regulasi. Insiden atau risiko tertentu harus diungkapkan kepada pengguna, regulator, atau masyarakat umum untuk memastikan akuntabilitas dan pengambilan keputusan yang tepat.

Persyaratan pengungkapan bervariasi menurut hukum dan konteks dan dapat mencakup:

- Penyebaran sistem AI berisiko tinggi
- Penerapan sistem AI berisiko tinggi
- Penurunan kinerja material
- Kebocoran data signifikan yang memengaruhi input atau output AI
- Output yang tidak akurat atau menyesatkan yang memengaruhi kesehatan atau keselamatan publik
- Perilaku sistem di luar kasus penggunaan yang dimaksudkan

Beberapa peraturan yang mewajibkan transparansi meliputi:

- **Undang-Undang AI Uni Eropa:** Mewajibkan penyebar sistem AI berisiko tinggi untuk memelihara dokumentasi teknis, instruksi penggunaan, keluhan pengguna, kinerja, risiko dan perbaikan, serta rencana pemantauan pasca-pemasaran
- **AIDA Kanada:** Mewajibkan pengungkapan publik tentang tujuan dan fungsi sistem AI berdampak tinggi (pada saat penulisan ini, undang-undang ini belum diberlakukan)
- **CPRA California dan panduan FTC AS:** Mewajibkan perusahaan untuk menjelaskan pengambilan keputusan otomatis dalam konteks tertentu
- **LGPD Brasil:** Memberikan hak kepada individu untuk penjelasan dan peninjauan keputusan yang didorong oleh AI

Pengungkapan mungkin perlu disesuaikan dengan kebutuhan dan hak setiap audiens, termasuk:

- **Pengguna dan pelanggan:** Penjelasan sederhana dan dapat ditindaklanjuti tentang apa yang terjadi dan apa yang sedang dilakukan
- **Regulator:** Dokumentasi teknis, garis waktu, penilaian dampak, dan tindakan perbaikan
- **Mitra dan vendor:** Implikasi tingkat sistem dan pembaruan antarmuka
- **Pemangku kepentingan internal:** Analisis akar penyebab, kebijakan yang direvisi, detail mitigasi, dan pembaruan pelatihan

Organisasi dapat proaktif dan merencanakan insiden, termasuk menyiapkan pengungkapan jauh sebelum insiden terjadi. Contohnya meliputi:

- Pernyataan publik dan FAQ yang menjelaskan insiden dan respons organisasi
- Grafik garis waktu insiden
- Penjelasan perilaku model dalam bahasa yang mudah dipahami
- Informasi kontak untuk pertanyaan

Misalnya, sebuah perusahaan media sosial secara publik mengungkapkan bahwa AI moderasi kontennya gagal mendeteksi dan memblokir kampanye ujaran kebencian yang tersebar luas, yang memicu kerusakan reputasi dan peningkatan pengawasan regulasi.

Kesiapan dan Kesiapan Organisasi

Membangun program respons insiden AI membutuhkan perencanaan dan koordinasi yang cermat, pelatihan yang menyeluruh, dan infrastruktur yang kuat. Komponen utama dari program tersebut meliputi:

- Rencana Respons Insiden AI (IRP) yang terdokumentasi sebagai dokumen terpisah atau (idealnya) bagian dari IRP TI, keamanan siber, dan privasi secara keseluruhan
- Buku panduan untuk jenis insiden terkait AI yang umum
- Pengujian dan simulasi skenario respons
- Respons terintegrasi dengan tim keamanan siber dan hukum
- Program pelatihan bagi karyawan tentang deteksi dan pelaporan insiden

Pelatihan untuk responden insiden sering dikombinasikan dengan latihan simulasi, di mana seorang moderator memandu responden insiden melalui skenario insiden yang sedang berkembang, memungkinkan mereka untuk mendiskusikan pendekatan mereka terhadap insiden dan prosedur respons mereka. Simulasi seperti ini membantu membangun "memori otot" sehingga responden lebih siap ketika insiden nyata terjadi.

Skenario simulasi potensial meliputi:

- ✓ Model generatif menghasilkan materi yang dilindungi hak cipta
- ✓ Model prediktif gagal karena fitur input yang bergeser
- ✓ Tim merah mengungkapkan kerentanan keamanan dalam model dasar
- ✓ Pemburu ancaman menemukan bahwa penyerang eksternal baru-baru ini telah membahayakan sistem AI
- ✓ Regulator meminta semua log dari insiden baru-baru ini

11.4 RINGKASAN

Manajemen dan jaminan risiko AI merupakan aspek mendasar dari tata kelola AI yang bertanggung jawab, yang membahas bagaimana risiko diidentifikasi, dievaluasi, dan dikelola sepanjang siklus hidup sistem AI. Bab ini dimulai dengan pembahasan tentang peramalan dan kategorisasi risiko, termasuk domain risiko baru yang melampaui kerangka kerja TI tradisional. Praktisi dan manajer risiko harus mempertimbangkan risiko spesifik AI, termasuk bias, penyimpangan, kurangnya transparansi, dan kerugian hilir, terutama ketika komponen pihak ketiga atau sumber terbuka digunakan. Memahami perjanjian vendor, ketentuan lisensi, pembatasan penggunaan data, dan klausul tanggung jawab sangat penting untuk memahami sepenuhnya sejauh mana paparan yang ditanggung organisasi ketika mengadopsi alat AI eksternal.

Peramalan juga memainkan peran penting, yang mengharuskan tim tata kelola untuk mengantisipasi penggunaan sekunder, penyalahgunaan, dan ancaman yang muncul melalui simulasi, pemantauan penyimpangan, dan pemetaan dampak pemangku kepentingan.

Sifat sistem AI menggarisbawahi pentingnya audit kinerja, pengujian tim merah yang bersifat antagonis, pemodelan ancaman, dan pengujian keamanan. Audit memvalidasi fungsionalitas dunia nyata dan mengungkap masalah yang mungkin tidak muncul di lingkungan pengembangan atau selama pengujian. Red teaming memungkinkan pengungkapan kerentanan secara proaktif dengan mensimulasikan penggunaan berbahaya atau manipulasi cepat, yang sangat relevan untuk model dasar dan LLM. Pemodelan ancaman memperkenalkan kekhawatiran khusus AI seperti pencurian model atau serangan peracunan, sementara pengujian keamanan membantu memastikan kerahasiaan dan integritas data pelatihan, bobot model, dan lingkungan penyebaran. Perburuan ancaman membawa pertempuran ke musuh dengan secara proaktif mencari indikator kompromi. Aktivitas-aktivitas ini dan lainnya mendukung program jaminan yang memungkinkan keterlacakan, deteksi, dan ketahanan. Mereka juga mendukung kepatuhan terhadap mandat peraturan, yang semakin membutuhkan dokumen dan catatan formal, termasuk catatan audit, log penggunaan, dan tinjauan pihak ketiga.

Ketika masalah muncul, praktik yang kuat dan terdokumentasi untuk manajemen dan pengungkapan insiden memastikan dampak minimal dan kepatuhan terhadap peraturan. Insiden AI dapat mencakup keluaran yang tidak terduga, halusinasi model, rekomendasi berbahaya, sistem yang berperilaku di luar batas yang dimaksudkan, dan berbagai insiden dalam infrastruktur yang mendasarinya yang dapat mengkompromikan integritas dan kerahasiaan sistem AI yang terpengaruh.

Mengelola peristiwa-peristiwa ini melibatkan deteksi yang jelas, triase, komunikasi, investigasi, penahanan, pemulihan, tinjauan pasca-insiden, dan dokumentasi. Akar penyebab insiden AI dapat berasal dari model yang rapuh, kualitas data yang buruk, pengujian yang tidak memadai, atau penyimpangan model. Respons insiden yang efektif membutuhkan kolaborasi di antara semua pemangku kepentingan. Dokumentasi memberikan umpan balik ke register risiko dan laporan kepatuhan, sementara pengungkapan yang direncanakan dengan baik mendorong transparansi dan keselarasan peraturan. Model yang berhadapan dengan publik, khususnya yang digunakan di domain berisiko tinggi, harus disertai dengan instruksi yang jelas, rencana pemantauan pasca-pasar, dan pengungkapan yang disesuaikan untuk audiens yang berbeda. Semua praktik ini harus dianggap sebagai lapisan tambahan dari program dan praktik yang ada, termasuk manajemen risiko, audit, dan respons insiden.

BAB 12
OPERASI AI YANG BERKELANJUTAN

12.1 MENGELOLA CATATAN BISNIS

Catatan bisnis menceritakan kisah operasi dan proses suatu organisasi. Catatan ini sangat penting dalam konteks sistem AI, untuk memastikan akuntabilitas organisasi, kepatuhan terhadap peraturan, dan peningkatan sistem yang berkelanjutan. Dokumentasi sistematis tentang insiden, masalah, risiko, keputusan, dan aktivitas pemantauan pasca-pasar memberikan dasar untuk pengelolaan sistem AI yang bertanggung jawab di seluruh aktivitas dalam pengembangan dan tata kelola AI yang dibahas di seluruh buku ini. Bagian ini membahas praktik-praktik penting untuk memelihara catatan-catatan ini dan struktur tata kelola yang mendukung integritas jangka panjangnya.

Dokumentasi dan Pelacakan Insiden

Catatan insiden sangat penting untuk operasi dan tata kelola AI, karena memungkinkan organisasi untuk mendokumentasikan peristiwa yang menyimpang dari perilaku sistem yang dimaksudkan, baik yang disebabkan oleh bug perangkat lunak, penyimpangan model, masalah data, atau interaksi pengguna yang tidak terduga. Manajemen insiden dibahas secara lengkap di Bab 11.

Jenis insiden yang harus dicatat secara detail meliputi:

- **Kegagalan kinerja model:** Penurunan akurasi atau keandalan
- **Pelanggaran etika/regulasi:** Keluaran yang bias, keputusan yang tidak adil, dan rekomendasi yang berbahaya
- **Pelanggaran keamanan:** Masukan yang merugikan, ekstraksi model, peracunan model, dan serangan injeksi cepat
- **Waktu henti sistem:** Gangguan layanan sistem AI, terutama yang disebabkan oleh komponen AI
- **Kasus penggunaan yang tidak terduga:** Pengguna menerapkan sistem AI dengan cara yang tidak dimaksudkan

Apa yang Dianggap Sebagai Insiden dalam AI?

Hasil pencarian yang berperingkat buruk mungkin hanya gangguan kecil dalam mesin rekomendasi konten, tetapi kesalahan klasifikasi dalam alat diagnostik perawatan kesehatan dapat mengancam jiwa. Mengklasifikasikan suatu peristiwa sebagai "insiden" bergantung pada tingkat keparahan dan konteksnya.

Sebagai contoh, sistem pelacakan pelamar kerja suatu organisasi menggunakan alat AI untuk menyaring kandidat pekerjaan. Organisasi tersebut menemukan bahwa sistem tersebut secara sistematis memberi peringkat lebih rendah kepada pelamar yang lebih tua, sehingga membahayakan kepatuhan terhadap Peraturan Daerah NYC 144, yang mewajibkan audit bias

tahunan terhadap sistem pelacakan pelamar. Catatan insiden mencakup log dari kueri yang terpengaruh, analisis distribusi data pelatihan, temuan tinjauan internal, dan catatan tentang upaya pelatihan ulang dengan data yang telah dikoreksi.

Pelaporan Masalah dan Manajemen Siklus Hidup

Masalah dan kejadian mungkin tidak memenuhi ambang batas insiden, tetapi tetap mencerminkan perilaku sistem yang tidak diinginkan, keluhan pengguna, atau kekurangan teknis yang memerlukan investigasi. Hal ini mewakili masalah yang perlu ditangani secara formal dan keputusan dibuat tentang langkah selanjutnya.

Contoh masalah meliputi:

- ❖ Ketidakkonsistenan UI/UX kecil
- ❖ Klasifikasi yang salah secara sporadis
- ❖ Umpan balik dari penguji QA internal
- ❖ Saran dan keluhan dari forum komunitas pengguna

Organisasi harus menerapkan sistem pelaporan masalah menggunakan alat tiket atau alur kerja. Siklus hidup suatu masalah biasanya meliputi:

1. **Penerimaan:** Pengajuan masalah melalui saluran yang ditentukan.
2. **Kategorisasi:** Pemberian tag berdasarkan tingkat keparahan, area, dan komponen AI.
3. **Penugasan:** Pengarahan ke tim yang bertanggung jawab.
4. **Eskalasi:** Pengarahan ke pihak lain ketika SLA dilanggar.
5. **Resolusi:** Implementasi perbaikan atau peningkatan.
6. **Penutupan:** Mendokumentasikan resolusi dan memberi tahu pemangku kepentingan.

Organisasi yang matang melacak masalah dalam pipeline CI/CD dan alur kerja pelatihan ulang mereka, memungkinkan peringatan otomatis ketika masalah berulang menandakan degradasi model.

Dokumentasi Risiko dan Pelacakan Mitigasi

Karena sistem AI tunduk pada risiko unik, mendokumentasikan risiko, potensi kerugian, keputusan penanganan, dan upaya mitigasi sangat penting untuk tata kelola risiko. Risiko terkait AI meliputi:

- Pergeseran model
- Pergeseran data
- Risiko bias dan keadilan
- Kerentanan keamanan
- Risiko hukum dan kepatuhan
- Kekhawatiran penyalahgunaan atau penggunaan ganda

Setiap risiko harus didokumentasikan dalam register risiko, idealnya dalam platform tata kelola AI atau alat GRC (tata kelola, risiko, dan kepatuhan). Catatan register risiko tipikal biasanya mencakup:

- Nama sistem atau aplikasi
- Deskripsi risiko
- Kasus penggunaan terkait
- Skor kemungkinan dan dampak

- Keputusan penanganan risiko, termasuk orang yang membuat keputusan
- Strategi mitigasi (jika berlaku)
- Risiko residual pasca-mitigasi
- Pemilik risiko dan jadwal peninjauan

Sebagai contoh, sistem aplikasi pinjaman berbasis AI bank mencakup entri risiko untuk bias rasial karena variabel proksi dalam data historis. Catatan tersebut terdiri dari langkah-langkah mitigasi (misalnya, audit keadilan dan teknik pasca-pemrosesan), serta tanggal peninjauan berkala untuk menilai kembali efektivitasnya.

Rencana Pemantauan Pasca-Pemasaran

Memantau sistem AI setelah penyebarannya, yang disebut sebagai pemantauan pasca-pemasaran, sangat penting untuk mengidentifikasi masalah dunia nyata yang mungkin terlewatkan oleh pengujian dan validasi internal sebelum penyebaran. Tujuan pemantauan pasca-pemasaran umumnya meliputi:

- Validasi akurasi, ketahanan, dan keadilan model yang berkelanjutan
- Deteksi penurunan kinerja karena berbagai penyebab, termasuk perubahan data dan lingkungan
- Pengumpulan umpan balik pengguna dan adaptasi perilaku
- Dukungan untuk pelatihan ulang iteratif, penilaian risiko, dan pembaruan tata kelola

Komponen dari rencana pemantauan pasca-pemasaran meliputi:

- **Pemantauan metrik kinerja:** Pengamatan berkelanjutan terhadap indikator kinerja utama (misalnya, akurasi, skor F1, dan latensi) di seluruh segmen pengguna
- **Audit bias dan keadilan:** Tinjauan terjadwal menggunakan paritas demografis, kesempatan yang sama, atau uji keadilan kontrafaktual
- **Deteksi pergeseran:** Memantau distribusi input dan output untuk pergeseran statistik menggunakan metode seperti Indeks Stabilitas Populasi atau uji Kolmogorov–Smirnov
- **Tinjauan aliran dan transfer data:** Tinjauan berkala terhadap transfer data ke dan dari model AI, sistem lain, dan ke dan dari organisasi itu sendiri. Perangkat pencegahan kehilangan data (DLP) dapat membantu mengotomatiskan pemantauan ini. Pemantauan perlu mencakup transfer data lintas batas untuk memastikan organisasi tetap mematuhi peraturan terkait.
- **Saluran umpan balik pengguna:** Mengintegrasikan mekanisme umpan balik ke dalam antarmuka pengguna memungkinkan pengajuan keluhan atau saran secara real-time.
- **Integrasi insiden dan masalah:** Memasukkan anomali yang terdeteksi ke dalam proses pelaporan insiden dan masalah serta perangkat untuk pelacakan dan penyelesaian.

Contoh peraturan pemantauan pasca-pasar meliputi:

- ✓ **Undang-Undang AI Uni Eropa:** Mewajibkan sistem AI berisiko tinggi untuk menerapkan mekanisme pemantauan pasca-pasar yang mencakup pencatatan, pelaporan, dan pengawasan manusia.
- ✓ **Undang-Undang Akuntabilitas Algoritma AS:** Undang-undang yang diusulkan ini menyerukan penilaian dampak dan evaluasi berkelanjutan terhadap sistem yang diterapkan.

Struktur Tata Kelola dan Akuntabilitas

Agar efektif, akuntabilitas untuk pencatatan bisnis harus didokumentasikan secara formal di semua peran dan fungsi dalam siklus hidup sistem AI.

Peran dan tanggung jawab tipikal meliputi:

- **Pengembang sistem:** Memelihara catatan desain sistem, asal usul data pelatihan, silsilah data, rencana pengujian, hasil pengujian, dan versi model.
- **Manajer produk:** Memastikan bahwa dokumentasi yang berhadapan dengan pengguna akurat dan mutakhir.
- **Petugas kepatuhan:** Memverifikasi bahwa pencatatan selaras dengan mandat hukum dan peraturan.
- **Pemilik risiko:** Memantau status risiko terbuka dan keputusan penanganan risiko serta mengawasi rencana mitigasi untuk memastikan manajemen risiko yang efektif.
- **Pemilik alur kerja dan proses:** Memelihara catatan tinjauan dan persetujuan pengembangan sistem AI.

Tabel 12.1 menunjukkan RACI tipikal untuk pencatatan AI.

Tabel 12.1
Contoh Matriks RACI Pencatatan Data AI

Tugas	Bertanggung jawab	Akuntabel	Dikonsultasikan	Diinformasikan
Desain sistem AI	Pengembang	CIO	CISO, CIO, manajer produk	Tim eksekutif
Dokumen insiden	DevOps	Pemimpin risiko	Tim kepatuhan	Manajer produk
Pertahankan risiko daftar	Pemilik risiko	CISO	Penasihat hukum	Tim eksekutif
Memperbarui rencana pemantauan	AI/ML insinyur	Kepala AI	Pimpinan QA	Semua pemangku kepentingan

Kebijakan Retensi dan Pemusnahan

Catatan yang terkait dengan pengembangan dan pengoperasian sistem AI harus dimasukkan dalam jadwal retensi catatan organisasi. Biasanya, catatan harus disimpan untuk jangka waktu yang ditentukan secara hukum, biasanya dari tiga hingga tujuh tahun, tergantung pada yurisdiksi dan hukum yang berlaku. Organisasi harus menentukan:

- **Periode retensi:** Berdasarkan jenis sistem dan tingkat risiko
- **Kontrol akses:** Pastikan hanya personel yang berwenang yang dapat mengakses catatan (sebagian besar catatan harus tidak dapat diubah untuk mencegah perusakan dan penghancuran)
- **Prosedur pengarsipan:** Simpan dalam format yang aman dan tahan terhadap perusakan untuk pelestarian jangka panjang
- **Protokol penghapusan:** Menghapus catatan di akhir masa pakainya, menggunakan metode yang sesuai dengan risiko dan sensitivitas data

Alat dan Teknologi untuk Manajemen Catatan Bisnis

Seringkali masuk akal untuk menggunakan alat bantu untuk mengelola catatan bisnis, terutama karena sistem AI modern menghasilkan sejumlah besar data operasional. Log utama meliputi:

- **Log prediksi:** Input dan output model untuk setiap kueri
- **Log kesalahan:** Kegagalan atau pengecualian selama inferensi atau pemrosesan data
- **Log penggunaan:** Frekuensi dan konteks akses sistem

Platform tata kelola AI yang sedang berkembang mulai menawarkan fitur pencatatan dan kepatuhan khusus AI, termasuk:

- Kontrol versi model dan silsilah data
- Asal usul dan silsilah data
- Integrasi register risiko dan kepatuhan
- Dasbor pemantauan pasca-penyebaran
- Alur kerja respons insiden

Contoh platform meliputi Fiddler AI, Arthur.ai, Credo AI, dan Microsoft Responsible AI Dashboard.

Pertimbangan Hukum dan Regulasi

Pencatatan yang berkaitan dengan pengembangan dan pengoperasian sistem AI sering kali didorong oleh hukum, peraturan, dan ketentuan dalam perjanjian hukum. Contoh kerangka kerja meliputi:

- **UU AI Uni Eropa:** Mewajibkan sistem berisiko tinggi untuk memelihara dokumentasi teknis terperinci, log, dan catatan pemantauan
- **UU AI Brasil:** Mewajibkan sistem berisiko tinggi untuk mendokumentasikan klasifikasi risiko, penilaian dampak, log operasional, proses tata kelola, insiden, pengujian, dan intervensi manusia (pada saat penulisan buku ini, undang-undang ini belum diberlakukan)
- **GDPR:** Mewajibkan organisasi untuk memberikan informasi yang bermakna tentang keputusan otomatis dan menjaga akuntabilitas, yang dalam praktiknya seringkali memerlukan pencatatan log.
- **NIST AI RMF:** Mendorong dokumentasi di semua tahapan untuk kepercayaan dan kemampuan audit.
- **ISO/IEC 42001:** Menekankan pencatatan log, pelaporan insiden, dan dokumentasi risiko untuk sistem manajemen AI.

Hukum sangat bervariasi di setiap yurisdiksi, termasuk:

- Di Kanada, Undang-Undang Perlindungan Informasi Pribadi dan Dokumen Elektronik (PIPEDA) mewajibkan akuntabilitas atas hasil pengambilan keputusan otomatis, yang meluas ke sistem AI.
- Di Brasil, Lei Geral de Proteção de Dados (LGPD) berkaitan dengan kemampuan audit AI dalam pengambilan keputusan pemrosesan data.
- Di Tiongkok, peraturan AI mewajibkan transparansi algoritma dan audit rutin terhadap perilaku sistem.

Integrasi dengan Praktik Organisasi yang Lebih Luas

Catatan bisnis yang terkait dengan sistem AI tidak boleh diisolasi, melainkan harus diintegrasikan ke dalam alur kerja dan repositori operasional, kepatuhan, dan manajemen risiko organisasi lainnya. Pencatatan data AI harus selaras dengan kebijakan tata kelola informasi perusahaan yang ada, termasuk klasifikasi dokumen, sistem manajemen konten perusahaan (ECM), dan jejak audit.

Audit internal dan eksternal harus dilakukan untuk memastikan bahwa:

- ✓ Catatan terkait AI lengkap dan akurat, termasuk proses tata kelola, tinjauan, dan keputusan
- ✓ Kontrol akses berbasis peran ditegakkan
- ✓ Tanggapan insiden tepat waktu dan terdokumentasi dengan baik

Kepatuhan seharusnya bukan satu-satunya pendorong untuk pencatatan. Sebaliknya, pencatatan membantu organisasi dalam beberapa hal, termasuk:

- **Tinjauan pasca-insiden dan laporan RCA:** Ini memberikan wawasan berharga untuk mencegah insiden di masa mendatang.
- **Analisis tren risiko:** Ini membantu mengidentifikasi pola dan masalah yang muncul.
- **Pengulangan masalah:** Ini menandakan area yang membutuhkan peningkatan proses.

Misalnya, sistem deteksi penipuan keuangan berbasis AI mengalami lonjakan positif palsu. Analisis akar penyebab mengungkapkan peristiwa pergeseran data, yang mendorong pengembangan protokol pelatihan ulang yang lebih baik dan pemantauan yang ditingkatkan untuk mendeteksi pola serupa di masa mendatang.

12.2 KOMUNIKASI EKSTERNAL

Komunikasi eksternal merupakan aspek penting dari tata kelola AI, karena organisasi terkadang harus berkomunikasi dengan regulator, pelanggan, dan pemangku kepentingan lainnya. Komunikasi yang efektif dan jelas membantu organisasi mempertahankan kepercayaan, mematuhi harapan regulasi, dan menginformasikan publik, pengguna, dan mitra tentang berbagai hal. Bagian ini menjelaskan pengembangan dan pengelolaan rencana komunikasi eksternal terstruktur yang mendorong transparansi, akuntabilitas, dan keterlibatan pemangku kepentingan di seluruh siklus hidup sistem AI.

Membangun Struktur Tata Kelola Komunikasi

Komunikasi eksternal menyampaikan nada, suara, dan isi pesan kepada berbagai pihak, termasuk publik. Hal ini membutuhkan perencanaan, disiplin, dan kematangan untuk menghindari penerbitan siaran pers ad hoc atau pernyataan reaktif. Struktur tata kelola komunikasi memastikan bahwa pesan konsisten, terotorisasi, dan selaras dengan kebijakan kepatuhan dan manajemen risiko yang lebih luas.

Organisasi harus menunjuk personel kunci untuk mengawasi komunikasi eksternal terkait AI, dan peran tersebut dapat mencakup:

- ❖ **Petugas informasi publik (PIO):** Perwakilan resmi organisasi di forum publik
- ❖ **Pemimpin media sosial:** Mengelola kehadiran media sosial dan menyampaikan komunikasi melalui media sosial

- **Pemimpin komunikasi AI:** Bertanggung jawab atas strategi pesan khusus AI
- **Penasihat hukum:** Meninjau komunikasi untuk kepatuhan, pertimbangan kekayaan intelektual, dan risiko tanggung jawab
- **Manajer produk atau insinyur:** Memberikan penjelasan teknis

Komunikasi eksternal organisasi harus diatur oleh kebijakan yang terintegrasi dengan manajemen risiko perusahaan, tata kelola data, privasi, protokol respons insiden, dan kewajiban pelaporan peraturan. Kebijakan ini membantu menentukan kapan, bagaimana, dan oleh siapa komunikasi eksternal harus terjadi, terutama selama skenario krisis yang melibatkan sistem AI. Dalam sebagian besar atau semua kasus, komunikasi kepada pihak eksternal mana pun harus disetujui secara formal oleh pemimpin yang ditunjuk.

Misalnya, sebuah perusahaan rintisan perawatan kesehatan menerapkan sistem AI diagnostik dan menugaskan seorang petugas medis untuk menjelaskan kemampuan dan keterbatasannya, memastikan bahwa pengguna menerima panduan yang tepat secara medis daripada jargon pemasaran.

Mengkomunikasikan Kemampuan dan Keterbatasan Sistem AI

Para pemangku kepentingan eksternal, termasuk regulator, pengguna, mitra, dan masyarakat umum, berhak mendapatkan informasi yang jelas, ringkas, dan mudah diakses tentang sistem AI mereka, cara kerjanya, dan apa yang seharusnya (dan tidak seharusnya) diandalkan dari sistem tersebut.

Pengungkapan harus mencakup hal-hal berikut:

- **Tujuan dan kasus penggunaan:** Apa yang dirancang untuk dilakukan oleh sistem
- **Keterbatasan dan risiko yang diketahui:** Batasan akurasi, potensi bias, ketergantungan data, atau peringatan keandalan
- **Keterlibatan manusia:** Bagaimana dan kapan pengawasan manusia digunakan
- **Sumber data dan perilaku model:** Deskripsi umum data pelatihan dan bagaimana keputusan dibuat
- **Pertanyaan dan permintaan subjek data:** Cara mendapatkan bantuan, mengajukan banding atas keputusan, dan memperbaiki kesalahan

Pengungkapan memiliki banyak bentuk dan dapat muncul dalam:

- Dokumentasi produk dan manual pengguna
- Perjanjian hukum dan pernyataan ketentuan penggunaan
- Pemberitahuan privasi dan ketentuan penggunaan
- Portal web atau kartu model yang merangkum karakteristik teknis
- FAQ publik atau penjelasan yang ditulis dalam bahasa yang mudah dipahami

Misalnya, produk asisten suara bertenaga AI suatu organisasi menerbitkan halaman transparansi yang menjelaskan interaksi yang dicatat, bagaimana model AI menafsirkan perintah, dan perintah mana yang memicu penyimpanan data.

Bahaya Mengklaim Kemampuan AI Secara Berlebihan

Melebih-lebihkan kemampuan sistem AI dapat menyebabkan pengawasan regulasi, hilangnya kepercayaan konsumen, atau bahkan tuntutan hukum. Tim komunikasi harus menyeimbangkan antusiasme dengan akurasi teknis dan hukum.

Kewajiban Komunikasi Regulasi dan Hukum

Beberapa yurisdiksi dan kerangka peraturan mengharuskan organisasi untuk secara proaktif atau reaktif mengungkapkan informasi tentang sistem AI mereka kepada pihak berwenang atau publik.

Dalam beberapa kasus, seperti di bawah Undang-Undang AI Uni Eropa, sistem AI berisiko tinggi harus didaftarkan dalam basis data publik, yang mencakup dokumentasi tentang tujuan penggunaannya, spesifikasi teknis, langkah-langkah mitigasi risiko, dan rencana pemantauan pasca-pemasaran.

Insiden, seperti hasil yang bias, penurunan kinerja, atau serangan yang merugikan, mungkin memerlukan komunikasi eksternal sebagaimana dipersyaratkan oleh:

- GDPR (Pasal C4): Memberitahukan subjek data tentang pelanggaran berisiko tinggi
- Undang-Undang AI Uni Eropa: Memberi tahu regulator tentang insiden serius dan tindakan korektif
- Hukum perlindungan konsumen: Memberi tahu pengguna tentang pelanggaran, pengungkapan yang tidak sah, cacat produk, atau hasil yang menyesatkan

Organisasi yang matang mengembangkan templat komunikasi yang telah disetujui sebelumnya yang sesuai dengan harapan peraturan sambil mempertahankan nada dan citra organisasi. Ini dapat memperlancar respons selama pengungkapan yang sensitif terhadap waktu.

Mengelola Hubungan Masyarakat dan Media

Persepsi publik memainkan peran penting dalam membentuk reputasi organisasi, termasuk bagaimana penerapan AI dipersepsikan. Bahkan sistem yang secara teknis mumpuni pun dapat kehilangan kepercayaan publik jika komunikasi ditangani dengan buruk atau orang-orang bingung. Hubungan media yang efektif harus proaktif, transparan, dan terinformasi.

Dalam hal hubungan mereka dengan pelanggan dan publik, organisasi harus cenderung pada komunikasi proaktif, termasuk:

- Siaran pers yang mengumumkan penerapan AI baru
- Postingan blog publik yang menjelaskan pilihan desain
- Presentasi konferensi atau white paper
- Melibatkan panel penasihat eksternal atau dewan etika
- Webinar dan seminar untuk menjelaskan penerapan AI

Pendekatan proaktif membantu menetapkan harapan, membangun kepercayaan, dan menunjukkan itikad baik.

Bahkan perusahaan yang paling terorganisir pun terkadang mengalami masalah dan kejadian. Ketika terjadi kegagalan sistem atau kontroversi publik, rencana komunikasi krisis

harus mengarahkan PIO dan pihak lain tentang cara berkomunikasi secara efektif dengan pesan yang tepat. Komponen kunci meliputi:

- ✓ **Tim respons krisis:** Biasanya dipimpin oleh PIO, ini adalah kelompok yang ditunjuk dan diberi wewenang untuk berbicara kepada pihak berwenang, pers, dan publik.
- ✓ **Pembingkaian pesan:** Berpusat pada transparansi, tanggung jawab, dan tindakan.
- ✓ **Kesiapan media:** Termasuk siaran pers yang telah ditulis sebelumnya dan jawaban atas pertanyaan yang diantisipasi.

Misalnya, setelah sistem pengenalan wajah salah mengidentifikasi orang kulit berwarna, sebuah perusahaan teknologi segera menghentikan penyebaran, melakukan audit, menerbitkan temuan audit, dan terlibat dengan kelompok hak-hak sipil di forum terbuka untuk membahas respons dan perbaikan. (Perhatikan bahwa beberapa kota di AS dan yurisdiksi lain telah melarang atau membatasi penggunaan pengenalan wajah.)

Komunikasi dan Transparansi yang Menghadap Pengguna

Audiens eksternal yang paling berpengaruh adalah pelanggan dan pengguna akhir. Organisasi harus memastikan bahwa semua komunikasi yang ditujukan kepada pengguna mendorong penggunaan yang terinformasi dan bertanggung jawab, harapan, dan keterlibatan yang penuh hormat dengan alat AI.

Sistem dan situs yang digunakan pengguna dan pelanggan harus menyertakan label yang jelas yang menunjukkan:

- ➢ Bahwa sistem tersebut digerakkan oleh AI (misalnya, "dihasilkan oleh AI")
- ➢ Identitas pengembang
- ➢ Sifat keluaran sistem (misalnya, saran vs. keputusan)
- ➢ Mekanisme untuk menolak atau mengajukan banding

Contoh pelabelan dan pengungkapan yang diwajibkan termasuk pelabelan deepfake dan konten sintetis yang diamanatkan oleh Undang-Undang AI Uni Eropa, dan Undang-Undang Peningkatan Transparansi Online (BOT) California, yang mengharuskan bot untuk mengidentifikasi diri selama interaksi online untuk tujuan komersial.

Pengguna harus dapat secara intuitif mengajukan pertanyaan dan menyuarakan kekhawatiran tentang perilaku AI. Pengguna harus mendapatkan akses ke pusat bantuan, opsi kontak untuk eskalasi, dan menerima tanggapan yang tepat waktu dan penuh hormat. Organisasi harus mencatat dan menganalisis umpan balik ini sebagai bagian dari upaya pemantauan pasca-pasar mereka.

Komunikasi dengan Mitra dan Vendor

Sebagian besar organisasi melakukan outsourcing sebagian besar infrastruktur TI mereka, yang berarti tata kelola AI tidak berhenti di batas organisasi. Vendor, pengembang, dan integrator pihak ketiga juga harus dilibatkan dalam rencana komunikasi eksternal.

Saat bekerja dengan pihak ketiga, organisasi harus menyematkan rencana komunikasi ke dalam:

- ❖ **Perjanjian vendor:** Mewajibkan mitra untuk melaporkan gangguan, insiden, dan perubahan kinerja sistem

- **Perjanjian tingkat layanan (SLA):** Menentukan jangka waktu dan ruang lingkup untuk pembaruan, pengungkapan, dan penyelesaian masalah
- **Protokol komunikasi bersama:** Mengklarifikasi siapa yang berbicara atas nama sistem jika terjadi kepemilikan bersama

Dalam konteks sumber terbuka dan kolaboratif, komunikasi dengan mitra dan vendor harus mendukung:

- Publikasi kerentanan, pelanggaran, cacat, dan masalah lainnya secara transparan
- Peta jalan untuk pembaruan sistem dan model
- Saluran umpan balik dari kontributor

Misalnya, seorang pengembang AI yang bekerja dengan lembaga akademis memposting semua pembaruan dan batasan model di GitHub dan memberi tahu kontributor melalui milis dan panggilan komunitas.

Perencanaan Komunikasi di Seluruh Siklus Hidup AI

Komunikasi eksternal tidak boleh diperlakukan sebagai peristiwa sekali waktu, tetapi sebagai bagian dari proses keseluruhan yang dipetakan di seluruh tahapan siklus hidup AI, termasuk:

- **Desain dan pengembangan:** Publikasi penelitian dan pemberitahuan penggunaan data
- **Validasi dan pengujian:** Pengungkapan akurasi dan hasil audit bias
- **Penyebaran:** Pernyataan tujuan, panduan pengguna, dan pengarahan pemangku kepentingan
- **Operasi:** Pembaruan pasca-pasar, peringatan insiden, dan FAQ
- **Penghentian:** Pemberitahuan penghentian, pengungkapan arsip, dan informasi penghentian pengguna

Mengukur Efektivitas Komunikasi

Organisasi yang matang mengukur proses dan metrik publik mereka untuk melakukan perbaikan pada operasional mereka. Untuk memastikan upaya komunikasi eksternal efektif, organisasi harus mengevaluasi:

- **Kejelasan:** Apakah pengguna dan pemangku kepentingan memahami pesan yang dimaksud
- **Jangkauan:** Apakah audiens yang tepat diberi informasi
- **Responsif:** Apakah pertanyaan dan masalah ditangani tepat waktu
- **Indikator kepercayaan:** Apakah kepercayaan publik meningkat atau menurun seiring waktu.

Ide untuk metrik komunikasi eksternal meliputi:

- Survei kepuasan pengguna
- Analisis sentimen media
- Volume dan waktu penyelesaian pertanyaan
- Umpan balik pemangku kepentingan selama konsultasi
- Analisis keterlibatan media sosial
- Metrik lalu lintas ke halaman web yang berisi konten khusus AI

12.3 PENGHENTIAN PENGGUNAAN SISTEM AI

Seperti sistem dan aplikasi informasi tradisional, sistem AI tidak digunakan selamanya. Seperti teknologi perusahaan lainnya, sistem AI pada akhirnya harus ditingkatkan, diganti, dinonaktifkan, atau ditarik karena usang, penurunan kinerja, masalah kepatuhan hukum, atau pergeseran bisnis strategis. Siklus hidup AI yang terkelola dengan baik harus mencakup kebijakan yang jelas dan proses yang dapat diulang untuk menghentikan penggunaan sistem AI. Bagian ini memandu organisasi dalam merencanakan, menerapkan, dan mengelola penonaktifan, penarikan, atau lokalisasi sistem AI yang terkontrol.

Penghentian penggunaan harus menjadi bagian standar dari proses siklus pengembangan sistem (SDLC) organisasi mana pun, bukan aktivitas yang dibuat secara mendadak. Sama seperti manajer produk perangkat lunak yang bertanggung jawab merencanakan peningkatan versi dan dukungan akhir masa pakai, pengembang AI harus menyematkan strategi penonaktifan ke dalam cetak biru arsitektur. Perencanaan mengurangi keputusan yang didorong oleh kepanikan ketika penghentian penggunaan menjadi perlu.

Memahami Kebutuhan Penghentian Penggunaan Sistem AI

Penghentian penggunaan adalah fungsi tata kelola penting yang memastikan komponen AI tidak kehilangan kegunaannya atau menimbulkan risiko yang berkelanjutan dan meningkat. Beberapa faktor dapat memicu kebutuhan untuk menonaktifkan atau melokalisasi model. Pemicu umum untuk penonaktifan sistem AI meliputi:

- ✓ **Perubahan peraturan:** Persyaratan hukum baru dapat membuat penggunaan berkelanjutan menjadi tidak sesuai, atau biaya untuk menjadi sesuai mungkin terlalu mahal.
- ✓ **Penurunan kinerja:** Sistem AI mungkin tidak berkinerja sebaik sistem yang lebih baru atau mengalami pergeseran model, data usang, atau penurunan akurasi.
- ✓ **Prioritas ulang bisnis:** Pergeseran dalam strategi produk atau penawaran layanan organisasi dapat membuat sistem tersebut tidak dibutuhkan.
- ✓ **Kekhawatiran keamanan atau etika:** Munculnya kerentanan yang diketahui atau dampak yang berbahaya dapat mendorong organisasi untuk menghentikan penggunaan sistem AI.
- ✓ **Penarikan model pihak ketiga:** Vendor AI eksternal mungkin telah menghentikan penggunaan model AI atau API-nya.
- ✓ **Persyaratan lokalisasi:** Sistem AI mungkin tidak dapat mematuhi kewajiban regional (misalnya, kedaulatan data).

Sebagai contoh, AI yang dilatih terutama pada bahasa Inggris AS ditemukan menghasilkan keluaran yang bias dalam penerapan internasional. Organisasi tersebut memutuskan untuk melokalisasi dan melatih ulang model terpisah untuk pasar Eropa (sebagian untuk mempermudah upaya kepatuhan GDPR) dan menghentikan versi global di yurisdiksi tersebut.

Membangun Kebijakan Penghentian Sistem AI

Penghentian sistem apa pun harus diatur oleh kebijakan. Kebijakan penghentian sistem menyediakan struktur formal untuk memandu kapan dan bagaimana sistem AI dihentikan.

Kebijakan tersebut harus selaras dengan kerangka kerja tata kelola AI yang lebih luas dan didokumentasikan dalam standar manajemen siklus hidup AI organisasi.
Elemen inti dari kebijakan penghentian meliputi:

- **Kriteria penghentian:** Tentukan ambang batas operasional, keuangan, strategis, etis, dan hukum yang memicu pertimbangan untuk penghentian. Kriteria ini dapat mencakup kegagalan KPI yang berkelanjutan, biaya pemeliharaan yang berlebihan, perubahan strategi bisnis, konflik peraturan yang terkonfirmasi, atau tuntutan pemangku kepentingan.
- **Proses persetujuan dan eskalasi:** Identifikasi pihak yang bertanggung jawab untuk memulai dan menyetujui tindakan penghentian. Eskalasi harus terjadi ketika risiko etis atau kerugian pelanggan teridentifikasi.
- **Persyaratan perencanaan pensiun:** Mencakup langkah-langkah wajib seperti pemberitahuan pemangku kepentingan, pencadangan sistem, pengarsipan data, dan dukungan transisi pengguna.
- **Dokumentasi dan jejak audit:** Membutuhkan catatan rinci tentang alasan, waktu, metode, pemangku kepentingan, dan proses yang terlibat dalam pensiun.
- **Tinjauan pasca-tindakan (AAR):** Mencakup analisis pelajaran yang dipetik untuk menginformasikan praktik pengembangan dan penerapan di masa mendatang.

Pensiun AI vs. Penonaktifan TI

Meskipun serupa dalam konsep, pensiun AI melibatkan lebih dari sekadar mematikan sistem. Pensiun sistem AI mungkin memerlukan penghapusan pelatihan, tinjauan etika, penghapusan model turunan, dan komunikasi publik, terutama ketika diwajibkan oleh peraturan atau ketika keputusan berdampak tinggi sebelumnya diotomatisasi.

Perencanaan untuk Deaktivasi Terkendali

Deaktivasi terkendali dari sistem AI harus dilakukan sebagai bagian dari proses tertulis yang berulang, untuk mengurangi gangguan, memitigasi risiko, dan melestarikan catatan penting. Langkah-langkah untuk deaktivasi terkendali meliputi:

1. **Pemetaan pemangku kepentingan:** Identifikasi pemangku kepentingan, termasuk pengguna internal, pelanggan eksternal, regulator, dan mitra. Tentukan kebutuhan mereka selama transisi.
2. **Analisis penggunaan:** Petakan ketergantungan untuk lebih memahami di mana sistem tertanam dan alat atau layanan hilir mana yang bergantung padanya (jika sistem AI merupakan subjek penilaian dampak bisnis—suatu aktivitas yang terkait dengan perencanaan pemulihan bencana—ketergantungan ini mungkin sudah diketahui).
3. **Perencanaan cadangan atau transisi:** Tentukan apakah model lain, proses manual, atau alat pihak ketiga akan mengambil alih semua atau sebagian fungsi yang dilakukan oleh sistem AI yang akan dihentikan.

4. **Pemberitahuan dan edukasi pengguna:** Berkomunikasi secara jelas dengan pihak yang terkena dampak dan berikan panduan tentang perubahan, prosedur baru, alasan penghentian, dan bagaimana mereka dapat mengajukan pertanyaan tambahan.
5. **Cuplikan audit akhir:** Tangkap keadaan akhir model AI, termasuk input, bobot, metrik kinerja, dan log keputusan, untuk memastikan akuntabilitas dan transparansi.
6. **Eksekusi penutupan:** Nonaktifkan API, hentikan titik akhir model, dan hapus titik integrasi. Perbarui alat keamanan seperti kontrol akses jaringan dan firewall sesuai kebutuhan.
7. **Pemantauan pasca-penonaktifan:** Pantau kesehatan sistem dan perilaku pengguna untuk memastikan transisi yang sukses.

Seringkali, ketika suatu sistem dinonaktifkan, sebagian atau seluruh otomatisasi diambil alih oleh sistem pengganti. Aktivitas keseluruhan dikenal sebagai migrasi, yang seringkali melibatkan migrasi data dari sistem lama ke sistem baru. Hal ini juga seringkali membutuhkan pengujian yang cukup banyak untuk memastikan bahwa data telah dimigrasikan dengan benar dan bahwa sistem pengganti berfungsi dengan benar. Seringkali, sistem yang diganti ditempatkan dalam mode "hanya baca" atau "pelaporan" jika organisasi diharuskan untuk menyediakan informasi untuk tujuan historis untuk periode tertentu.

Penghentian penggunaan sistem AI dapat menimbulkan tantangan operasional dan tata kelola, termasuk:

- **Kelangsungan bisnis:** Memastikan bahwa proses-proses penting yang sebelumnya didukung oleh sistem AI terus didukung dan tetap beroperasi.
- **Integritas data:** Mencegah hilangnya jejak audit atau ketertelusuran.
- **Pencabutan akses:** Menghapus kunci API, kredensial, dan akun layanan.
- **Manajemen risiko residual:** Memantau konsekuensi yang tidak diinginkan dari penghentian, seperti penurunan pengalaman pengguna atau kesalahan proses manual.

Misalnya, sebuah bank menghentikan penggunaan model penilaian risiko pinjaman berbasis AI setelah menemukan pergeseran demografi pelamar yang tidak lagi diwakili oleh model tersebut. Pelanggan diberitahu dua bulan sebelumnya, dan penjamin emisi manusia untuk sementara menangani aplikasi sementara model pengganti diperoleh, dilatih, dan divalidasi.

Strategi Lokalisasi dan Penghentian Penggunaan Sistem AI di Tingkat Regional

Terkadang, model tidak dihentikan penggunaannya secara global, tetapi dinonaktifkan atau dilokalisasi untuk wilayah tertentu, karena perbedaan lingkungan kepatuhan atau konteks budaya. Terkadang terlalu sulit untuk membuat satu sistem AI mematuhi berbagai peraturan dan harapan pelanggan.

Faktor pendorong lokalisasi dapat meliputi:

- **Hukum kedaulatan data:** Mandat yang mengharuskan pelatihan atau inferensi terjadi di dalam perbatasan nasional (misalnya, Tiongkok atau Rusia).
- **Norma etika dan sosial:** Sensitivitas regional dapat memerlukan penyesuaian dalam praktik penyaringan konten atau rekomendasi.

- **Kendala peraturan:** Undang-undang seperti Undang-Undang AI Uni Eropa dapat membatasi atau melarang penggunaan kasus penggunaan sistem AI tertentu yang dianggap tidak dapat diterima atau berisiko tinggi.

Pendekatan untuk melakukan lokalisasi meliputi:

- **Instans model terpisah:** Menerapkan model khusus negara yang dilatih pada data yang dilokalisasi.
- **Pembelajaran terfederasi:** Menggunakan pelatihan terdesentralisasi untuk mempertahankan residensi data lokal.
- **Arsitektur modular:** Memungkinkan pengaktifan selektif fitur atau konten regional.

Sebagai contoh, sebuah perusahaan media global menjalankan AI moderasi konten di berbagai benua. Karena hukum kebebasan berbicara regional, perusahaan tersebut mempertahankan seperangkat aturan dan model yang dilatih ulang yang berbeda di Prancis, Jerman, dan Amerika Serikat, dan menonaktifkan versi global untuk penerapan khusus yurisdiksi.

Artefak Data dan Model: Pengarsipan dan Pembuangan

Ketika suatu sistem dinonaktifkan, organisasi harus menentukan informasi apa yang harus disimpan, diarsipkan, atau dihapus. Peraturan, perjanjian hukum, atau kebijakan internal mengatur keputusan ini.

Informasi yang mungkin dipilih oleh suatu organisasi untuk diarsipkan meliputi:

- Versi model final dan silsilah data pelatihan
- Log kinerja dan laporan kesalahan
- Laporan insiden dan catatan mitigasi
- Output keputusan, terutama untuk kasus penggunaan berdampak tinggi (misalnya, perawatan kesehatan dan keuangan)
- Catatan komunikasi pengguna dan pemangku kepentingan
- Dokumen desain dan artefak lain yang menjelaskan sistem AI
- Informasi lain apa pun yang ditentukan oleh peraturan sebagai informasi yang wajib disimpan

Informasi yang mungkin dipilih oleh suatu organisasi untuk dibuang meliputi:

- Data pelatihan yang berlebihan atau usang
- Artefak model sementara (misalnya, file titik pemeriksaan sementara)
- Informasi identitas pribadi apa pun yang tidak diperlukan untuk kepatuhan

Organisasi terkadang menggunakan alat penghapusan data tingkat militer untuk membuang informasi sensitif secara permanen, memastikan bahwa informasi tersebut tidak dapat dipulihkan.

Organisasi yang menggunakan sistem AI yang dihosting oleh penyedia layanan cloud perlu meminta penyedia layanan cloud untuk menghancurkan informasi tersebut, mungkin termasuk sertifikat penghancuran yang mungkin dibutuhkan organisasi untuk audit. Demikian pula, suatu organisasi mungkin perlu mencari solusi AI yang dihosting alternatif ketika penyedia AI yang dihosting saat ini mengumumkan penghentian sistemnya sendiri.

Komunikasi dan Transparansi dalam Penghentian AI

Seringkali, organisasi perlu memberi tahu pihak eksternal tentang niat mereka untuk menghentikan penggunaan sistem AI. Komunikasi yang ringkas sangat penting untuk memastikan bahwa semua pihak diberitahu tentang keputusan penghentian dan dampaknya. Contoh komunikasi ini meliputi:

- ✓ **Pemberitahuan pengguna:** Peringatan dalam produk, kampanye email, atau pembaruan dukungan
- ✓ **Pemberitahuan organisasi:** Komunikasi resmi kepada pelanggan dan mitra bisnis mungkin diperlukan oleh perjanjian hukum, termasuk syarat dan ketentuan yang berlaku, seperti waktu pemberitahuan
- ✓ **Pengungkapan publik:** Siaran pers, pengumuman media sosial, dan pernyataan di situs web organisasi atau portal transparansi
- ✓ **Pelaporan peraturan:** Pemberitahuan kepada pihak berwenang bila diwajibkan oleh hukum (misalnya, Undang-Undang AI Uni Eropa atau undang-undang perlindungan konsumen)
- ✓ **Pengarahan internal:** Pemberitahuan kepada tim hukum, kepatuhan, dukungan, dan penjualan

Tidak semua orang akan senang dengan pengumuman tentang penghentian sistem yang akan datang. Organisasi harus bersiap menghadapi:

- ➢ Penolakan dari pengguna yang bergantung pada sistem
- ➢ Kesalahpahaman tentang alasan penghentian
- ➢ Pertanyaan mengenai migrasi ke penyedia layanan lain
- ➢ Pengawasan regulasi terhadap waktu dan alasan

Keterlibatan awal tim komunikasi krisis dan PIO akan membantu organisasi mengantisipasi dan menanggapi pertanyaan-pertanyaan ini. Memiliki panduan komunikasi krisis yang dipersiapkan dengan baik dapat mengurangi risiko reputasi.

Misalnya, sebuah perusahaan teknologi menghentikan penggunaan alat gambar AI generatif karena kekhawatiran tentang penyalahgunaan. Perusahaan tersebut mengeluarkan siaran pers, memperbarui situs web AI yang bertanggung jawab, dan menawarkan pengembalian dana kepada pengguna atau jalur migrasi ke alat alternatif.

Penghentian sistem AI harus ditangani dengan cara yang konsisten dengan hukum dan kewajiban kontraktual di mana sistem tersebut diterapkan. Kewajiban hukum yang berlaku meliputi:

- ❖ **UU AI Uni Eropa:** Mewajibkan pemantauan berkelanjutan dan penarikan sistem berisiko tinggi yang tidak patuh
- ❖ **GDPR:** Mewajibkan minimalisasi data dan mungkin mengharuskan penghapusan data pribadi
- ❖ **Hukum perlindungan konsumen:** Dapat memicu kewajiban untuk mengembalikan dana atau memberikan kompensasi kepada pengguna yang terkena dampak penghapusan fitur berbasis AI

- **Ketentuan kontrak:** Mungkin memerlukan periode pemberitahuan atau jaminan ketersediaan

Risiko Litigasi Selama Penghentian Operasi

Penghentian operasi sistem AI yang tidak tepat atau tiba-tiba dapat membuat organisasi rentan terhadap tuntutan hukum, terutama jika pelanggan bergantung pada keluaran sistem, atau jika dampak diskriminatif ditemukan setelah penghentian operasi (alasan lain untuk mengarsipkan informasi sistem).

Pelajaran yang Dipetik dan Siklus Umpan Balik Organisasi

Penghentian sistem AI seharusnya tidak hanya berfungsi sebagai peristiwa akhir masa pakai, tetapi juga sebagai sumber pembelajaran dan peningkatan untuk penerapan di masa mendatang. Organisasi yang matang memanfaatkan peluang untuk merefleksikan, belajar, dan meningkatkan melalui tinjauan pasca-tindakan (AAR).

Praktik tinjauan pasca-penghentian penggunaan yang penting meliputi:

- Analisis akar penyebab: Mengapa sistem perlu dihentikan penggunaannya
- Penilaian dampak: Pengaruhnya terhadap pengguna, operasional, dan reputasi bisnis
- Tinjauan etika pasca-penghentian penggunaan: Untuk memastikan bahwa model AI yang telah dihentikan penggunaannya tidak digunakan kembali dengan cara yang merugikan oleh pihak lain
- Evaluasi kebijakan: Apakah proses penghentian penggunaan selaras dengan kebijakan formal
- Pengumpulan pengetahuan: Informasi apa yang harus didokumentasikan untuk tim di masa mendatang

12.4 RINGKASAN

Operasi AI yang berkelanjutan membutuhkan disiplin dan pendekatan terstruktur untuk mengelola catatan, berkomunikasi dengan pemangku kepentingan, dan menonaktifkan sistem secara bertanggung jawab. Seiring AI terus berkembang dan meluas di berbagai industri, tata kelola tidak berakhir pada saat penerapan; tata kelola berlanjut sepanjang masa operasional sistem dan seterusnya, termasuk saat penghentiannya.

Pencatatan merupakan bagian penting dari tata kelola AI. Dengan mendokumentasikan insiden, masalah, risiko, dan aktivitas pemantauan pasca-pasar, organisasi menciptakan jejak audit yang mendukung akuntabilitas, pembelajaran, dan kepatuhan terhadap peraturan. Proses ini melibatkan lebih dari sekadar pelaporan reaktif: ini mencakup pelacakan proaktif terhadap masalah kecil, register risiko terstruktur, dan strategi pemantauan untuk mendeteksi penyimpangan model atau kerugian yang tidak disengaja di lapangan. Alat-alat seperti log insiden, pelacak masalah, dan dasbor kinerja memungkinkan tim untuk menanggapi perkembangan dunia nyata dan terus meningkatkan sistem AI mereka.

Komunikasi eksternal juga penting untuk tata kelola AI. Organisasi harus menjaga transparansi dengan pengguna, regulator, mitra, dan publik. Ini termasuk pengungkapan yang

jelas tentang apa yang dilakukan sistem AI, bagaimana cara kerjanya, dan di mana keterbatasannya berada. Baik melalui dokumentasi produk, kartu model publik, atau pemberitahuan insiden, komunikasi eksternal harus akurat, mudah diakses, dan selaras dengan persyaratan hukum. Ekspektasi peraturan, termasuk yang ada dalam Undang-Undang AI Uni Eropa atau GDPR, seringkali mewajibkan komunikasi ini, sementara kepercayaan publik bergantung padanya. Strategi media yang proaktif, peran dan tanggung jawab yang didefinisikan dengan jelas, dan rencana komunikasi krisis lebih lanjut memastikan bahwa organisasi dapat merespons secara efektif dan tepat dalam situasi apa pun.

Pada akhirnya, setiap sistem AI mencapai titik di mana pengoperasian berkelanjutan tidak lagi layak. Perencanaan pensiun merupakan bagian penting dari siklus hidup AI, pensiun berdasarkan desain, yang memungkinkan penonaktifan atau lokalisasi model yang bertanggung jawab sebagai respons terhadap perubahan peraturan, penurunan kinerja, dan masalah etika. Organisasi harus menerapkan kebijakan dan praktik pensiun terstruktur yang mendefinisikan kapan dan bagaimana sistem harus dimatikan, bagaimana pengguna dan regulator harus diberitahu, dan data apa yang harus diarsipkan atau dihapus. Penonaktifan terkontrol, strategi cadangan, penyesuaian regional, dan tinjauan pembelajaran membantu memastikan bahwa penghentian sistem berjalan tertib dan berisiko rendah.

DAFTAR PUSTAKA

Akpan, E. E., & Essien, F. G. (2025). Artificial Intelligence: A Dependable Tool for Effective Management of Resources in Secondary Schools. *Shared Seasoned International Journal of Topical Issues*, *11*(1), 119-130.

Batool, A., Zowghi, D., & Bano, M. (2025). AI governance: a systematic literature review. *AI and Ethics*, *5*(3), 3265-3279.

Bélisle-Pipon, J. C., Monteferrante, E., Roy, M. C., & Couture, V. (2023). Artificial intelligence ethics has a black box problem. *AI & society*, *38*(4), 1507-1522.

Birkstedt, T., Minkkinen, M., Tandon, A., & Mäntymäki, M. (2023). AI governance: themes, knowledge gaps and future agendas. *Internet Research*, *33*(7), 133-167.

Burgess, A. (2018). The executive guide to artificial intelligence. *How to identify and implement applications for AI in your organization. London: AJBurgess Ltd*, *3*.

Carter, D. (2020). Regulation and ethics in artificial intelligence and machine learning technologies: where are we now? Who is responsible? Can the information professional play a role?. *Business Information Review*, *37*(2), 60-68.

Cath, C. (2018). Governing artificial intelligence: ethical, legal and technical opportunities and challenges. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2133).

Chhillar, D., & Aguilera, R. V. (2022). An eye for artificial intelligence: Insights into the governance of artificial intelligence and vision for future research. *Business & Society*, *61*(5), 1197-1241.

Cihon, P., Schuett, J., & Baum, S. D. (2021). Corporate governance of artificial intelligence in the public interest. *Information*, *12*(7), 275.

Daly, A., Hagendorff, T., Hui, L., Mann, M., Marda, V., Wagner, B., ... & Witteborn, S. (2019). Artificial intelligence governance and ethics: global perspectives. *arXiv preprint arXiv:1907.03848*.

De Almeida, P. G. R., Dos Santos, C. D., & Farias, J. S. (2021). Artificial intelligence regulation: a framework for governance. *Ethics and Information Technology*, *23*(3), 505-525.

Djeffal, C. (2019). Artificial intelligence and public governance: normative guidelines for artificial intelligence in government and public administration. In *Regulating artificial intelligence* (pp. 277-293). Cham: Springer International Publishing.

Dunleavy, P., & Margetts, H. (2025). Data science, artificial intelligence and the third wave of digital era governance. *Public Policy and Administration*, *40*(2), 185-214.

El Arab, R. A., Al Moosa, O. A., Sagbakken, M., Ghannam, A., Abuadas, F. H., Somerville, J., & Al Mutair, A. (2025). Integrative review of artificial intelligence applications in nursing: education, clinical practice, workload management, and professional perceptions. *Frontiers in Public Health*, *13*, 1619378.

Gillan, C., Hodges, B., Wiljer, D., & Dobrow, M. (2021). Health care professional association agency in preparing for artificial intelligence: Protocol for a multi-case study. *JMIR Research Protocols*, *10*(5), e27340.

González-Espejo, M. J., & Pavón, J. (2020). An introductory guide to artificial intelligence for legal professionals.

Goos, M., & Savona, M. (2024). The governance of artificial intelligence: Harnessing opportunities and mitigating challenges. *Research Policy*, *53*(3), 104928.

Guidance, W. H. O. (2021). Ethics and governance of artificial intelligence for health. *World Health Organization*, 1-165.

Hakim, A., & Anggraini, P. (2023). Artificial intelligence in teaching Islamic studies: Challenges and opportunities. *Molang: Journal Islamic Education*, *1*(2), 19-30.

Hohma, E., Boch, A., Trauth, R., & Lütge, C. (2023). Investigating accountability for artificial intelligence through risk governance: A workshop-based exploratory study. *Frontiers in Psychology*, *14*, 1073686.

Janssen, M., Brous, P., Estevez, E., Barbosa, L. S., & Janowski, T. (2020). Data governance: Organizing data for trustworthy Artificial Intelligence. *Government information quarterly*, *37*(3), 101493.

Jefferson-Jones, J. (2018). Advising the Smart City: When Artificial Intelligence and Big Data Are the Subjects of Professional Advice, What Is a Local Government Lawyer to Do. *U. Tol. L. Rev.*, *50*, 447.

Jia, Q., Guo, Y., Li, R., Li, Y., & Chen, Y. (2018). A conceptual artificial intelligence application framework in human resource management.

Johnson, B. A., Coggburn, J. D., & Llorens, J. J. (2022). Artificial intelligence and public human resource management: questions for research and practice. *Public Personnel Management*, *51*(4), 538-562.

Kavitha, V., & Lohani, R. (2019). A critical study on the use of artificial intelligence, e-Learning technology and tools to enhance the learners experience. *Cluster Computing*, *22*(Suppl 3), 6985-6989.

Kelley, S. (2026). Developing an artificial intelligence ethics governance checklist for the legal community. *AI and Ethics*, *6*(1), 47.

Kerr, A., Barry, M., & Kelleher, J. D. (2020). Expectations of artificial intelligence and the performativity of ethics: Implications for communication governance. *Big Data & Society*, *7*(1), 2053951720915939.

Khare, K., Stewart, B., & Khare, A. (2018). Artificial intelligence and the student experience: An institutional perspective. The International Academic Forum (IAFOR).

Larsson, S. (2020). On the governance of artificial intelligence through ethics guidelines. *Asian Journal of Law and Society*, *7*(3), 437-451.

Lauterbach, A. (2019). Artificial intelligence and policy: quo vadis?. *Digital Policy, Regulation and Governance*, *21*(3), 238-263.

Leenen, J. P., Hiemstra, P., Ten Hoeve, M. M., Jansen, A. C., van Dijk, J. D., Vendel, B., ... & Hettinga, M. (2025). Exploring the complex nature of implementation of Artificial intelligence in clinical practice: an interview study with healthcare professionals, researchers and policy and governance experts. *PLOS Digital Health*, *4*(5), e0000847.

Leslie, D. (2019). Understanding artificial intelligence ethics and safety. *arXiv preprint arXiv:1906.05684*.

Limna, P., Jakwatanatham, S., Siripipattanakul, S., Kaewpuang, P., & Sriboonruang, P. (2022). A review of artificial intelligence (AI) in education during the digital era. *Advance Knowledge for Executives*, *1*(1), 1-9.

Liua, Y., Salehb, S., & Huang, J. (2021). Artificial intelligence in promoting teaching and learning transformation in schools. *Artificial Intelligence*, *15*(3), 1-12.

Marchant, G. E., & Gutierrez, C. I. (2022). Soft law 2.0: an agile and effective governance approach for artificial intelligence. *Minn. JL Sci. & Tech.*, *24*, 375.

Mariam, G., Adil, L., & Zakaria, B. (2024). The integration of artificial intelligence (ai) into education systems and its impact on the governance of higher education institutions. *International Journal of Professional Business Review: Int. J. Prof. Bus. Rev.*, *9*(12), 13.

Mishra, A. K., Tyagi, A. K., Dananjayan, S., Rajavat, A., Rawat, H., & Rawat, A. (2024). Revolutionizing government operations: The impact of artificial intelligence in public administration. *Conversational Artificial Intelligence*, 607-634.

Morley, J., Murphy, L., Mishra, A., Joshi, I., & Karpathakis, K. (2022). Governing data and artificial intelligence for health care: developing an international understanding. *JMIR formative research*, *6*(1), e31623.

Müller, R., Locatelli, G., Holzmann, V., Nilsson, M., & Sagay, T. (2024). Artificial intelligence and project management: Empirical overview, state of the art, and guidelines for future research. *Project Management Journal*, *55*(1), 9-15.

Pertuit, T. L. (2026). Leveraging Artificial Intelligence to Enhance Comprehensive School Counseling Programs: An ASCA–MTSS–AI Conceptual Framework. *Professional School Counseling*, *30*(1), 2156759X261421920.

Qian, Y., Siau, K. L., & Nah, F. F. (2024). Societal impacts of artificial intelligence: Ethical, legal, and governance issues. *Societal impacts*, *3*, 100040.

Radu, R. (2021). Steering the governance of artificial intelligence: national strategies in perspective. *Policy and society*, *40*(2), 178-193.

Renda, A. (2019). *Artificial Intelligence. Ethics, governance and policy challenges*. CEPS Centre for European Policy Studies.

Ryan, M., & Stahl, B. C. (2021). Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications. *Journal of Information, Communication and Ethics in Society*, *19*(1), 61-86.

Salutina, T. Y., Platunina, G. P., & Frank, I. A. (2024, November). The Role of Artificial Intelligence in Improving the Effectiveness of Professional Training of Students of Electrical Engineering and Electronic Engineering. In *2024 Intelligent Technologies and Electronic Devices in Vehicle and Road Transport Complex (TIRVED)* (pp. 1-5). IEEE.

Smuha, N. A. (2019). The EU approach to ethics guidelines for trustworthy artificial intelligence. *Computer Law Review International*, *20*(4), 97-106.

Tapalova, O., & Zhiyenbayeva, N. (2022). Artificial intelligence in education: AIEd for personalised learning pathways. *Electronic Journal of e-Learning*, *20*(5), 639-653.

Thierer, A. D. (2023). Flexible, pro-innovation governance strategies for artificial intelligence. *R Street Policy Study*, (283).

Wirtz, B. W., Weyerer, J. C., & Kehl, I. (2022). Governance of artificial intelligence: A risk and guideline-based integrative framework. *Government information quarterly*, *39*(4), 101685.

World Health Organization. (2021). *Ethics and governance of artificial intelligence for health: WHO guidance*. World Health Organization.

Zhang, J., & Zhang, Z. M. (2023). Ethics and governance of trustworthy medical artificial intelligence. *BMC medical informatics and decision making*, *23*(1), 7.

TATA KELOLA PROFESIONAL dengan AI

(Kecerdasan Buatan)

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

BIO DATA PENULIS

Penulis memiliki berbagai disiplin ilmu yang diperoleh dari Universitas Diponegoro (UNDIP) Semarang. dan dari Universitas Kristen Satya Wacana (UKSW) Salatiga. Disiplin ilmu itu antara lain teknik elektro, komputer, manajemen dan ilmu sosiologi. Penulis memiliki pengalaman kerja pada industri elektronik dan sertifikasi keahlian dalam bidang Jaringan Internet, Telekomunikasi, Artificial Intelligence, Internet Of Things (IoT), Augmented Reality (AR), Technopreneurship, Internet Marketing dan bidang pengolahan dan analisa data (komputer statistik).

Penulis adalah pendiri dari Universitas Sains dan Teknologi Komputer (Universitas STEKOM) dan juga seorang dosen yang memiliki Jabatan Fungsional Akademik Lektor Kepala (Associate Professor) yang telah menghasilkan puluhan Buku Ajar ber ISBN, HAKI dari beberapa karya cipta dan Hak Paten pada produk IPTEK. Sejak tahun 2023 penulis tercatat sebagai Dosen luar biasa di Fakultas Ekonomi & Bisnis (FEB) Universitas Diponegoro Semarang. Penulis juga terlibat dalam berbagai organisasi profesi dan industri yang terkait dengan dunia usaha dan industri, khususnya dalam pengembangan sumber daya manusia yang unggul untuk memenuhi kebutuhan dunia kerja secara nyata.

YAYASAN PRIMA AGUS TEKNIK

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

ISBN 978-634-7227-90-4 (PDF)
9 786347 227904