

BAB 5. Keamanan, Etika, Risiko, dan Tata Kelola AI

5.1 Pendahuluan

Pada bab-bab sebelumnya, pembaca telah diajak memahami AI sebagai bagian dari transformasi digital sektor publik, mengenali jenis-jenis solusi AI, menilai use case yang layak, lalu merancang dan mengintegrasikan solusi tersebut ke dalam proses bisnis, aplikasi organisasi, dan arsitektur teknis yang lebih nyata. Sampai pada titik itu, pembaca telah melihat bahwa AI dapat memberi banyak manfaat bila digunakan secara tepat: membantu membaca informasi yang besar, mempercepat orientasi awal, mendukung klasifikasi, memperkuat pengelolaan pengetahuan, membantu penyusunan draft awal, serta meningkatkan efisiensi pada berbagai proses kerja. Namun, justru karena AI semakin dekat dengan proses kerja yang nyata, maka kebutuhan akan **pagar pengaman** juga menjadi semakin besar. Di sinilah Bab 5 mengambil peran penting. Bab ini membahas keamanan, etika, risiko, dan tata kelola AI sebagai dasar yang harus menyertai seluruh pemanfaatan AI di organisasi.

Bab ini penting karena dalam praktik, pembahasan AI sering terlalu cepat bergerak ke manfaat dan inovasi, sementara aspek keamanan, etika, dan tata kelola ditempatkan belakangan. Padahal, untuk sektor publik, pendekatan seperti itu tidak cukup. Pemerintahan tidak hanya bekerja untuk mencapai efisiensi, tetapi juga untuk menjaga akuntabilitas, keadilan, perlindungan data, keamanan sistem, integritas informasi, serta kepercayaan masyarakat. Karena itu, pertanyaan tentang AI tidak boleh berhenti pada “apa yang bisa dilakukan AI”, tetapi harus dilanjutkan dengan pertanyaan: **apa risikonya, bagaimana mengendalikannya, nilai apa yang harus dijaga, dan tata kelola seperti apa yang perlu dibangun?** Prinsip-prinsip dalam Peta Jalan Kecerdasan Artifisial Nasional Tahun 2026–2029 sendiri menempatkan kebermanfaatan, keadilan, akuntabilitas, perlindungan data pribadi, transparansi, keamanan, keberlanjutan, integritas, inklusivitas, dan keterlibatan serta pengawasan manusia sebagai landasan pelaksanaan AI.

Kebutuhan untuk menempatkan keamanan dan tata kelola sebagai bagian inti AI juga ditegaskan dalam **Guidelines for Secure AI System Development**. Dokumen itu menyatakan bahwa AI memang memiliki potensi manfaat yang besar, tetapi manfaat tersebut hanya dapat diwujudkan bila sistem AI dikembangkan, dideploy, dan dioperasikan secara aman serta bertanggung jawab. Dokumen yang sama menegaskan bahwa keamanan siber merupakan prasyarat bagi keselamatan, ketahanan, privasi, fairness, efektivitas, dan keandalan sistem AI, dan bahwa keamanan harus menjadi persyaratan inti sepanjang siklus hidup sistem, bukan hanya pada fase pengembangan. Selain itu, pedoman tersebut juga menegaskan bahwa sistem AI menghadapi kerentanan keamanan yang baru di samping ancaman keamanan siber yang sudah lazim pada sistem biasa.

Pentingnya kerangka etika juga ditegaskan dalam **Konsep Pedoman Etika Kecerdasan Artifisial**. Dokumen tersebut menjelaskan bahwa perkembangan AI membawa tantangan

etika yang kompleks, seperti bias algoritma, pelanggaran penggunaan data, dan disinformasi yang berdampak pada individu, masyarakat, dan negara. Karena itu, pengembangan, penyelenggaraan, dan pemanfaatan AI perlu dilakukan secara bertanggung jawab, adil, transparan, menghormati privasi, menjamin keamanan, dan memungkinkan pengawasan manusia. Pedoman etika itu juga memuat unsur-unsur yang sangat relevan untuk tata kelola AI, yaitu nilai etika, manfaat penerapan etika, identifikasi dan mitigasi risiko, langkah-langkah perlindungan, peran dan tanggung jawab para pihak, serta pemantauan dan evaluasi.

Dari sudut pandang pembelajaran, Bab 5 dapat dipahami sebagai **pagar utama** untuk semua bidang. Baik AI dipakai untuk kesehatan, hukum, kebijakan publik, pengawasan internal, layanan administrasi, komunikasi publik, maupun pengetahuan internal, pertanyaan tentang keamanan, etika, risiko, dan tata kelola tetap harus muncul. Hal ini penting karena bidang yang berbeda boleh jadi memiliki use case yang berbeda, tetapi banyak risiko dasarnya saling beririsan. Data tetap bisa bocor. Model tetap bisa menghasilkan keluaran yang keliru. Prompt bisa dimanipulasi. Vendor bisa menjadi titik risiko. Pengguna bisa terlalu percaya pada hasil AI. Dan organisasi bisa jatuh pada otomatisasi berlebihan bila tidak menetapkan batas yang jelas. Karena itu, bab ini tidak ditujukan hanya untuk pembaca teknis, melainkan untuk semua peserta lintas disiplin yang terlibat dalam pemanfaatan AI.

Dalam konteks yang lebih konseptual, **ISO/IEC 22989** juga memberi dasar penting untuk membaca AI secara hati-hati. Standar tersebut menegaskan bahwa AI system adalah engineered system yang menghasilkan keluaran seperti konten, prediksi, rekomendasi, atau keputusan untuk tujuan yang ditentukan manusia, dan bahwa AI systems do not understand; mereka memerlukan pilihan desain, rekayasa, dan pengawasan manusia. Standar yang sama juga menegaskan bahwa siklus hidup AI berbeda dari sistem non-AI biasa karena lebih iteratif, cenderung kurang dapat diprediksi, dan memerlukan pengalaman operasional serta penyesuaian agar tujuan dapat tercapai. Selain itu, standar tersebut menekankan pendekatan manajemen risiko yang sistematis, berkelanjutan, dan mencakup privasi, keamanan digital, keselamatan, serta bias di setiap fase siklus hidup AI.

Pemahaman seperti ini penting karena ia menggeser cara pandang organisasi. AI tidak boleh diperlakukan seperti fitur biasa yang cukup dipasang lalu digunakan. AI perlu diperlakukan sebagai sistem yang:

- mempunyai risiko khusus,
- memerlukan pengawasan terus-menerus,
- memerlukan dokumentasi dan keterlacakan,
- serta membutuhkan keseimbangan antara inovasi dan kontrol.

Jika organisasi mengabaikan hal tersebut, maka AI yang semula dimaksudkan untuk membantu justru dapat:

- memperbesar kerentanan keamanan,
- menyebarkan informasi yang salah,
- melanggar privasi,
- memperkuat bias,
- atau mengaburkan tanggung jawab manusia.

Dari sinilah dapat dipahami mengapa keamanan AI tidak bisa disamakan begitu saja dengan keamanan sistem biasa. Sistem AI tidak hanya membawa risiko pada tingkat infrastruktur, aplikasi, dan data, tetapi juga pada tingkat model, prompt, vendor, dan keluaran. AI juga membuka ruang risiko yang lebih khas, seperti:

- keluaran yang tampak meyakinkan tetapi keliru;
- penggunaan model pihak ketiga yang kurang transparan;
- perubahan perilaku sistem karena perubahan data atau prompt;
- kebocoran informasi melalui interaksi model;
- sampai kemungkinan agentic AI menjalankan langkah yang tidak diinginkan bila batasnya tidak dirancang dengan baik. Pedoman keamanan AI menekankan bahwa organisasi harus memperlakukan model, data, prompt, software, dokumentasi, dan log sebagai aset yang perlu diidentifikasi, dilindungi, dilacak, dan dikendalikan, serta menerapkan kontrol atas data yang dapat diakses sistem AI dan terhadap konten yang dihasilkan AI sesuai tingkat sensitivitasnya.

Pada saat yang sama, bab ini juga penting karena keamanan AI tidak dapat dipisahkan dari etika dan tata kelola. Dalam banyak organisasi, keamanan kadang dipahami semata-mata sebagai urusan teknis: melindungi server, API, atau akses sistem. Padahal, dalam sektor publik, AI yang aman tetapi tidak adil, tidak dapat dijelaskan, atau tidak menghormati privasi tetap merupakan masalah. Sebaliknya, AI yang etis tetapi tidak aman juga bermasalah, karena kerentanannya dapat membuka jalan bagi manipulasi, kebocoran data, atau kerusakan kepercayaan. Itu sebabnya Bab 5 menyatukan empat tema besar—keamanan, etika, risiko, dan tata kelola—karena keempatnya memang saling terkait dan tidak sebaiknya dipisahkan secara tajam.

Secara lebih khusus, bab ini akan membahas beberapa lapisan risiko dan pengaman. Pertama, pembaca akan diajak memahami **mengapa keamanan AI berbeda dari keamanan sistem biasa**, agar sejak awal terlihat bahwa AI membawa kerentanan tambahan yang tidak boleh diabaikan. Kedua, bab ini akan membahas **risiko pada data, model, output, prompt, dan vendor**, karena semua unsur tersebut dapat menjadi sumber masalah bila tidak dikendalikan. Ketiga, pembahasan akan masuk ke **privasi dan perlindungan data**, serta **bias, fairness, dan integritas informasi**, karena dua hal ini merupakan isu yang sangat penting di sektor publik. Keempat, bab ini akan mengulas **risiko khusus generative AI dan agentic AI**, termasuk confabulation, prompt injection, data leakage, over-automation, dan tindakan agent yang tidak terkendali. Kelima, pembaca akan dibawa ke pembahasan **prinsip etika AI dan nilai publik, human oversight dan safeguard, manajemen risiko AI**, serta **tata kelola AI dalam organisasi pemerintah**, termasuk kebutuhan akan kebijakan internal dan tata kelola vendor. Struktur ini sengaja dibuat agar pembaca bergerak dari pemahaman risiko teknis dan operasional menuju pemahaman etika, pengawasan, dan pengaturan kelembagaan yang lebih luas.

Penting juga untuk dipahami bahwa bab ini bukan ditulis untuk menakut-nakuti pembaca terhadap AI. Tujuan bab ini justru sebaliknya: membantu pembaca melihat AI secara lebih matang. Organisasi yang baik bukan organisasi yang menolak AI hanya karena ada risiko, dan juga bukan organisasi yang menerima AI tanpa pagar yang cukup. Organisasi yang matang adalah organisasi yang:

- memahami risiko,
- menilai konteks penggunaannya,
- menyiapkan safeguard,
- menetapkan peran dan tanggung jawab,
- serta terus memantau dan memperbaiki sistemnya dari waktu ke waktu.

Pendekatan seperti ini sejalan dengan Konsep Pedoman Etika Kecerdasan Artifisial, yang menekankan bahwa pemangku kebijakan dan pelaku sektor perlu melaksanakan manajemen risiko di setiap tahap siklus hidup AI, memantau dampaknya secara berkala, menjamin transparansi dan akuntabilitas, serta menyiapkan mekanisme penilaian mandiri, pelaporan insiden, dan sistem pemantauan terpadu. Dokumen tersebut juga menggarisbawahi bahwa kategori risiko tinggi, misalnya yang melibatkan data pribadi, data biometrik, atau penggunaan yang berdampak langsung pada publik, memerlukan persyaratan dan pengawasan khusus.

Dari sudut pandang pemimpin digital dan pengelola organisasi, pelajaran utama dari pendahuluan ini adalah bahwa AI tidak boleh hanya dibaca sebagai peluang inovasi, tetapi juga sebagai tanggung jawab kelembagaan. Ketika sebuah instansi memutuskan memakai AI, instansi tersebut pada saat yang sama memutuskan untuk:

- mengelola risiko baru,
- menyiapkan kontrol baru,
- memperjelas akuntabilitas baru,
- dan membangun bentuk pengawasan baru yang sebelumnya mungkin belum dibutuhkan pada sistem biasa.

Dengan kata lain, semakin serius organisasi ingin memakai AI, semakin serius pula organisasi harus membangun pagar keamanan, etika, risiko, dan tata kelola di sekelilingnya.

Melalui dasar pemahaman ini, pembaca diharapkan siap memasuki pembahasan yang lebih rinci pada subbagian-subbagian berikutnya. Dari sini, diskusi akan bergerak ke pertanyaan pertama yang sangat penting, yaitu **mengapa keamanan AI berbeda dari keamanan sistem biasa**, agar pembaca dapat melihat bahwa AI membawa sifat, kerentanan, dan kebutuhan pengamanan yang memang memerlukan perhatian khusus sejak tahap desain sampai operasi.

5.2 Mengapa keamanan AI berbeda dari keamanan sistem biasa

Ketika organisasi mulai menggunakan AI, sering muncul anggapan bahwa pengamanannya pada dasarnya sama saja dengan sistem digital lain. Selama server aman, akses dibatasi, aplikasi diuji, dan jaringan dijaga, maka sistem dianggap cukup terlindungi. Cara pandang ini hanya sebagian benar. Memang, solusi AI tetap mewarisi banyak kebutuhan keamanan yang sudah dikenal dalam sistem informasi biasa, seperti:

- kontrol akses,
- keamanan jaringan,
- keamanan aplikasi,
- pengelolaan akun,
- pencatatan log,
- perlindungan data,
- dan penanganan kerentanan perangkat lunak.

Namun, AI membawa **lapisan risiko tambahan** yang membuat keamanannya tidak dapat dipahami hanya sebagai perpanjangan dari keamanan sistem biasa. Perbedaan ini penting, terutama dalam sektor publik, karena organisasi pemerintah tidak hanya mengelola aplikasi dan data, tetapi juga mengelola pengetahuan, layanan, keputusan awal, dan proses kerja yang semakin bergantung pada keluaran AI. Ketika AI masuk ke dalam proses kerja, maka yang harus diamankan bukan hanya infrastruktur dan aplikasi, tetapi juga:

- data yang dipakai AI,
- model atau engine yang menghasilkan keluaran,
- prompt atau instruksi yang mempengaruhi perilaku sistem,
- integrasi dengan pihak ketiga,
- dan keluaran AI itu sendiri.

Pedoman **Guidelines for Secure AI System Development** menegaskan secara sangat jelas bahwa sistem AI menghadapi **kerentanan keamanan yang baru**, di samping ancaman siber yang sudah lazim pada sistem tradisional. Dokumen tersebut juga menekankan bahwa keamanan siber adalah dasar bagi keselamatan, ketahanan, privasi, fairness, dan keandalan AI, serta harus diterapkan sepanjang siklus hidup sistem, bukan hanya pada tahap pengembangan.

Dengan demikian, pertanyaan yang harus dijawab organisasi bukan hanya “apakah sistem AI ini aman seperti aplikasi lain”, tetapi juga “apa karakter AI yang membuat permukaan serangannya berbeda, apa aset baru yang perlu dilindungi, dan bagaimana perilaku sistem dapat berubah dengan cara yang tidak biasa bagi sistem tradisional”.

Pada sistem digital biasa, organisasi umumnya berpikir tentang aset seperti:

- server,
- basis data,
- aplikasi,
- kode sumber,

- akun pengguna,
- dan koneksi jaringan.

Pada sistem AI, daftar aset ini menjadi lebih luas. Pedoman keamanan AI menyebut bahwa organisasi perlu memperlakukan **model, data, prompt, log, dokumentasi, dan komponen perangkat lunak terkait AI** sebagai aset yang perlu diidentifikasi, dilindungi, ditelusuri, dan dikelola.

Ini berarti solusi AI memiliki beberapa aset tambahan yang pada sistem biasa mungkin belum terlalu menonjol, misalnya:

- **model** atau engine inferensi;
- **training data** atau sumber pengetahuan;
- **prompt** dan *system prompt*;
- **retrieval corpus** atau basis dokumen;
- **parameter konfigurasi** yang mempengaruhi perilaku keluaran;
- serta **feedback loop** dari penggunaan sehari-hari.

Masing-masing aset itu dapat menjadi titik risiko. Jika model berubah, perilaku sistem bisa berubah. Jika corpus dokumen salah atau tercemar, jawaban AI bisa ikut salah. Jika prompt sistem dimanipulasi, keluaran AI dapat menyimpang dari batas yang seharusnya. Ini berbeda dari sistem biasa yang logika perilakunya cenderung lebih tetap dan lebih eksplisit.

5.2.1 Perilaku sistem AI

Salah satu perbedaan mendasar AI dibanding sistem biasa adalah bahwa perilaku AI cenderung **kurang sepenuhnya dapat diprediksi**. Pada sistem tradisional berbasis aturan yang tegas, alur keputusan biasanya lebih eksplisit: jika kondisi A terpenuhi, maka hasil B dihasilkan. Pada AI, terutama machine learning dan generative AI, hubungan antara input dan output tidak selalu sesederhana itu.

ISO/IEC 22989 menjelaskan bahwa siklus hidup AI cenderung lebih iteratif dan kurang dapat diprediksi dibanding sistem non-AI biasa, serta memerlukan pengalaman operasional dan penyesuaian berkelanjutan. Standar ini juga menegaskan bahwa AI systems do not understand dan bahwa output AI bersifat *error-prone*.

Dari sudut keamanan, hal ini penting karena:

- keluaran yang salah tidak selalu muncul dalam pola yang mudah ditebak;
- perubahan kecil pada input dapat menghasilkan keluaran yang sangat berbeda;
- sistem dapat tampak baik pada pengujian awal tetapi berperilaku berbeda pada kondisi operasional;
- dan pengguna dapat sulit membedakan kapan sistem sedang “tepat” dan kapan sedang “meyakinkan tetapi keliru”.

Dalam sistem biasa, keamanan sering banyak bertumpu pada mencegah akses tidak sah atau mencegah aplikasi rusak. Pada AI, organisasi juga perlu memikirkan **bagaimana perilaku sistem bisa dimanipulasi, digeser, atau dipicu menghasilkan keluaran yang tidak semestinya**, bahkan tanpa harus merusak infrastruktur utamanya.

Sistem AI memperluas apa yang dapat disebut sebagai **permukaan serangan**. Pada sistem biasa, permukaan serangan umumnya dipikirkan dalam bentuk endpoint, API, akun, database, atau jaringan. Pada AI, permukaan serangan mencakup hal-hal tambahan seperti:

- data masukan yang mempengaruhi inferensi;
- prompt yang diberikan pengguna;
- dokumen yang dimasukkan ke basis pengetahuan;
- model pihak ketiga;
- pipeline data;
- dan integrasi tool pada agentic AI.

Ini berarti penyerang atau penyalahgunaan tidak selalu harus “masuk ke server” untuk menimbulkan masalah. Pada beberapa kasus, cukup dengan:

- memberi input yang dirancang untuk menyesatkan model,
- memanipulasi prompt,
- memasukkan konten berbahaya ke corpus dokumen,
- atau mengeksploitasi hubungan antar-tool dalam workflow agentic.

Pedoman keamanan AI menekankan bahwa organisasi harus mengendalikan data yang dapat diakses AI, membatasi tindakan yang dapat dilakukan AI, menerapkan sanitasi input, dan membangun guardrails yang cukup terhadap interaksi model.

Dengan kata lain, pada AI, yang harus diamankan bukan hanya “pintu masuk ke aplikasi”, tetapi juga **jalur pengaruh terhadap perilaku model**.

Dalam banyak sistem biasa, input dipahami sebagai data yang akan diproses sesuai logika aplikasi. Pada AI generatif dan agentic AI, input sering sekaligus berfungsi sebagai **instruksi**. Inilah yang membuat keamanan AI berbeda secara nyata.

Sebuah prompt tidak hanya mengisi data ke sistem, tetapi dapat:

- mengarahkan fokus model,
- mempengaruhi cara sistem membaca sumber,
- mendorong keluaran tertentu,
- atau bahkan berusaha menabrak batas yang telah ditetapkan.

Karena itu, input pada AI perlu diperlakukan lebih hati-hati. Sistem AI tidak hanya membaca “apa isi masukan”, tetapi juga berusaha menafsirkan “apa yang diminta pengguna”. Hal ini membuka risiko yang pada sistem biasa lebih jarang muncul dalam bentuk serupa, seperti:

- prompt injection,
- penyelundupan instruksi tersembunyi,
- atau usaha membuat sistem mengabaikan aturan awalnya.

Inilah salah satu alasan mengapa keamanan AI tidak cukup hanya mengandalkan validasi input tradisional. Organisasi juga perlu memikirkan bagaimana sistem menangani instruksi yang menyesatkan, manipulatif, atau bertentangan dengan fungsi yang seharusnya.

5.2.2 Pembentukan Perilaku Data pada AI

Pada sistem informasi biasa, data sering dipandang sebagai sesuatu yang disimpan, ditampilkan, atau diproses. Pada AI, data memiliki posisi yang lebih dalam: data bisa **membentuk perilaku sistem**. Ini berlaku baik pada:

- training data,
- reference data,
- retrieval corpus,
- log umpan balik,
- maupun data yang dipakai saat inferensi.

Jika data yang masuk bermasalah, maka sistem tidak hanya menampilkan data yang salah, tetapi bisa menghasilkan perilaku yang salah. Misalnya:

- training data yang bias dapat membuat model bias;
- dokumen yang tidak mutakhir dapat membuat jawaban AI usang;
- corpus yang tercemar dapat membuat knowledge assistant memberi rujukan yang salah;
- dan feedback yang salah arah dapat memperburuk perilaku sistem dari waktu ke waktu.

Pedoman keamanan AI menekankan perlunya pengelolaan aset data secara hati-hati, termasuk sumber, ruang lingkup, keterbatasan, failure modes, retensi, dan frekuensi peninjauan.

Dengan demikian, keamanan AI menuntut organisasi melihat data bukan hanya sebagai objek yang dilindungi, tetapi juga sebagai unsur yang mempengaruhi bagaimana sistem berpikir dan bertindak.

Pada sistem biasa, perubahan perilaku sistem umumnya datang dari perubahan kode atau konfigurasi aplikasi. Pada AI, perilaku bisa berubah karena lebih banyak hal, misalnya:

- model diperbarui;
- prompt sistem diubah;
- basis dokumen diganti;
- parameter inference disesuaikan;
- atau sumber eksternal yang dipanggil agent berubah.

Ini berarti kontrol perubahan pada AI harus lebih disiplin. Organisasi perlu mengetahui:

- perubahan apa yang dilakukan;
- siapa yang mengubah;
- mengapa perubahan dilakukan;
- dan bagaimana dampaknya terhadap hasil sistem.

Pedoman keamanan AI secara khusus menekankan bahwa perubahan pada model, data, atau prompt dapat mengubah perilaku sistem, sehingga perubahan tersebut harus dikelola secara hati-hati, termasuk melalui dokumentasi dan versioning.

Perbedaan ini penting karena pada sistem biasa, perubahan bisa lebih mudah ditelusuri ke file konfigurasi atau modul kode. Pada AI, perubahan yang tampak kecil—misalnya revisi prompt atau penambahan dokumen baru ke knowledge base—bisa menghasilkan perubahan keluaran yang cukup besar.

Pada sistem biasa, keluaran aplikasi biasanya mengikuti logika yang lebih tetap. Pada AI, **output** itu sendiri menjadi sumber risiko yang perlu dikelola. Ini salah satu pembeda penting.

Keluaran AI bisa:

- meyakinkan tetapi salah;
- terlalu luas sehingga melampaui kewenangan;
- membocorkan informasi yang seharusnya tidak ditampilkan;
- merangkum secara menyesatkan;
- atau memicu tindakan manusia berdasarkan informasi yang lemah.

Karena itu, keamanan AI tidak hanya soal menjaga agar “sistem tidak diserang”, tetapi juga soal menjaga agar **keluaran sistem tidak menjadi pintu masuk kerugian**. Dalam sektor publik, hal ini sangat penting karena keluaran AI bisa mempengaruhi:

- pelayanan,
- komunikasi internal,
- orientasi kebijakan,
- tindak lanjut pengawasan,
- atau penyusunan dokumen administratif.

Inilah sebabnya *human-in-the-loop*, logging, monitoring, dan pembatasan penggunaan hasil menjadi bagian integral dari keamanan AI, bukan sekadar bagian etika atau operasional.

Pada banyak solusi AI modern, organisasi memanfaatkan:

- model pihak ketiga,
- API eksternal,
- platform cloud,
- atau layanan vendor yang sudah menyediakan komponen AI tertentu.

Ketergantungan ini bisa juga terjadi pada sistem biasa, tetapi pada AI implikasinya sering lebih luas karena vendor tidak hanya menyediakan infrastruktur, melainkan bisa juga mempengaruhi:

- perilaku model,
- pembaruan prompt sistem,
- kebijakan data,

- hingga kualitas output.

NIST AI 600-1 menekankan pentingnya inventarisasi pihak ketiga yang memiliki akses ke konten organisasi, evaluasi risiko pemasok, pencatatan perubahan, due diligence atas privasi dan keamanan, serta rencana kontinjensi bila layanan pihak ketiga bermasalah.

Ini menunjukkan bahwa keamanan AI lebih sensitif terhadap persoalan vendor dibanding banyak sistem biasa. Organisasi tidak cukup menilai “apakah vendor ini tepercaya”, tetapi juga harus menilai:

- data apa yang keluar ke vendor,
- model atau layanan apa yang dipakai,
- perubahan apa yang bisa dilakukan pihak ketiga,
- dan bagaimana organisasi tetap menjaga kendali serta akuntabilitasnya.

5.2.3 Pengamanan terhadap AI

Sistem biasa memang juga memerlukan pengamanan sepanjang siklus hidup, tetapi pada AI hal ini menjadi lebih menonjol karena perubahan data, model, dan konteks penggunaan bisa terus mempengaruhi sistem. Pedoman keamanan AI menegaskan bahwa keamanan harus diterapkan **di sepanjang seluruh siklus hidup AI**, bukan hanya di fase desain atau deployment.

Artinya, keamanan AI perlu dipikirkan pada:

- pemilihan dan pengelolaan data;
- pembangunan atau pemilihan model;
- desain prompt dan guardrails;
- integrasi ke aplikasi dan workflow;
- pengujian;
- operasi dan monitoring;
- pembaruan komponen;
- sampai penghentian atau penggantian sistem.

Pendekatan ini penting karena sistem AI bisa tampak aman saat awal, tetapi menjadi berisiko ketika:

- sumber dokumen berubah,
- model diperbarui,
- pengguna menggunakan sistem di luar tujuan awal,
- atau workflow agentic diperluas tanpa kontrol tambahan.

Akhirnya, perbedaan keamanan AI dari keamanan sistem biasa juga terletak pada fakta bahwa keamanan AI lebih sulit dipisahkan dari **etika dan tata kelola**. Pada sistem biasa, sering ada kecenderungan memisahkan keamanan sebagai isu teknis dan etika sebagai isu kebijakan. Pada AI, pemisahan ini jauh lebih sulit.

Sebagai contoh:

- kebocoran data bukan hanya masalah teknis, tetapi juga masalah privasi dan kepercayaan;
- bias model bukan hanya masalah kualitas, tetapi juga fairness;
- output yang menyesatkan bukan hanya masalah sistem, tetapi juga integritas informasi;
- over-automation bukan hanya masalah desain workflow, tetapi juga akuntabilitas;
- dan kegagalan guardrail bukan hanya masalah teknis, tetapi juga tata kelola.

Keamanan AI berbeda dari keamanan sistem biasa karena AI membawa aset baru yang harus dilindungi, perilaku yang lebih sulit diprediksi sepenuhnya, permukaan serangan yang lebih luas, input yang sekaligus berfungsi sebagai instruksi, data yang membentuk perilaku sistem, output yang sendiri bisa menjadi sumber risiko, serta ketergantungan yang lebih sensitif pada model dan vendor pihak ketiga. Karena itu, organisasi tidak cukup menerapkan pola keamanan tradisional secara

apa adanya, tetapi perlu memperluas cara pandanganya terhadap apa yang harus diamankan dan bagaimana pengamanan itu dilakukan sepanjang siklus hidup AI.

5.3 Risiko pada data, model, output, prompt, dan vendor

Setelah pada subbagian sebelumnya dijelaskan bahwa keamanan AI berbeda dari keamanan sistem biasa, langkah berikutnya adalah memetakan **di mana risiko-risiko utama itu berada**. Pemetaan ini penting karena dalam praktik organisasi, risiko AI sering dibicarakan secara terlalu umum. Orang mengatakan bahwa AI “berisiko”, “bisa salah”, atau “perlu dijaga”, tetapi tanpa kejelasan tentang objek apa yang sebenarnya sedang dibicarakan. Akibatnya, organisasi sulit menentukan kontrol yang tepat.

Dalam solusi AI, risiko tidak berada hanya pada satu titik. Risiko dapat muncul pada:

- **data** yang menjadi pondasi sistem,
- **model** atau engine yang menghasilkan keluaran,
- **output** yang dipakai manusia atau sistem lain,
- **prompt** yang mengarahkan perilaku AI,
- dan **vendor atau pihak ketiga** yang menyediakan komponen penting.

Cara membaca seperti ini membantu organisasi berpindah dari rasa khawatir yang umum menuju penilaian yang lebih operasional. Setiap titik memiliki karakter risiko yang berbeda, dan karena itu memerlukan bentuk pengendalian yang berbeda pula.

Pendekatan ini sejalan dengan **Guidelines for Secure AI System Development**, yang menekankan bahwa organisasi perlu mengidentifikasi dan melindungi aset AI seperti model, data, prompt, log, dokumentasi, dan komponen perangkat lunak terkait AI, serta menerapkan kontrol yang tepat atas data yang dapat diakses sistem AI dan tindakan yang dapat dipicu sistem. Dokumen tersebut juga menegaskan bahwa keamanan AI harus mencakup desain, pengembangan, deployment, serta operasi dan pemeliharaan.

ISO/IEC 22989 juga memberi dasar yang serupa. Standar tersebut menjelaskan bahwa AI system life cycle menuntut perhatian pada tujuan, requirement, risk management, operation and monitoring, continuous validation, dan re-evaluation, serta bahwa AI systems cenderung lebih iteratif dan kurang dapat diprediksi dibanding sistem non-AI biasa. Dengan demikian, memetakan risiko pada data, model, output, prompt, dan vendor bukan hanya latihan teknis, tetapi bagian dari tata kelola siklus hidup AI.

5.3.1 Risiko pada data

Risiko pertama dan seringkali paling mendasar adalah **risiko pada data**. Hal ini wajar, karena data bukan sekadar bahan baku AI. Data juga membentuk perilaku sistem, mempengaruhi kualitas keluaran, dan menentukan apakah solusi AI layak dipercaya.

Risiko pada data dapat muncul dalam beberapa bentuk.

Pertama, data yang salah, tidak lengkap, tidak konsisten, atau tidak mutakhir. Bila AI dibangun atau dijalankan di atas data seperti ini, maka hasilnya juga akan rentan salah. Dalam konteks sektor publik, hal ini bisa terjadi pada:

- metadata dokumen yang tidak seragam,
- label historis yang tidak konsisten,
- dokumen kebijakan yang belum diperbarui,
- atau data layanan yang banyak kosong.

Kedua, bias pada data. Data yang kurang mewakili variasi kondisi nyata dapat membuat AI bekerja lebih baik untuk sebagian kasus dan lebih buruk untuk kasus lain. Ini berpotensi menghasilkan perlakuan yang tidak adil, penandaan yang tidak seimbang, atau ringkasan yang mengabaikan perspektif tertentu.

Ketiga, data poisoning atau pencemaran data. Pedoman keamanan AI menekankan pentingnya sanitasi data dan input, termasuk ketika organisasi menggabungkan umpan balik pengguna atau data continuous learning, karena training data ikut menentukan perilaku sistem. Jika data yang masuk tercemar, dimanipulasi, atau tidak cukup dikendalikan, maka perilaku sistem bisa bergeser ke arah yang tidak diinginkan.

Keempat, risiko pada data sensitif dan data pribadi. Dalam sektor publik, banyak data berkaitan dengan identitas, layanan, dokumen internal, atau informasi yang tidak seharusnya tersebar. Bila data tersebut digunakan AI tanpa kontrol yang tepat, risikonya tidak hanya bersifat teknis, tetapi juga hukum, etika, dan reputasional.

Kelima, risiko pada provenance atau asal-usul data. Organisasi perlu tahu data berasal dari mana, siapa pemiliknya, bagaimana kualitasnya, dan dalam konteks apa data itu layak dipakai. Pedoman keamanan AI menekankan pentingnya dokumentasi sumber data, ruang lingkup, keterbatasan, serta failure modes untuk mendukung transparansi dan akuntabilitas.

Dari semua itu, satu pelajaran utama dapat diambil: risiko data bukan sekadar risiko “database bocor”. Risiko data dalam AI juga berarti risiko bahwa **sistem belajar, membaca, atau bertindak di atas pondasi yang salah**.

5.3.2 Risiko pada model

Risiko kedua berada pada **model** atau engine inferensi yang menjadi pusat perilaku AI. Pada sistem biasa, logika program cenderung lebih eksplisit. Pada AI, terutama machine learning dan generative AI, model bekerja dengan pola statistik atau representasi yang lebih sulit dibaca secara langsung. Karena itu, model membawa risiko yang khas.

Pertama, ada risiko bahwa **model tidak cukup cocok untuk use case**. Pedoman keamanan AI mengingatkan bahwa pemilihan model perlu mempertimbangkan kesesuaian model dengan use case, kemampuan menjelaskan output, karakteristik training dataset, dan trade-off keamanan lainnya. Dokumen tersebut bahkan menegaskan bahwa model yang lebih sederhana dan lebih transparan kadang lebih baik daripada model yang besar tetapi lebih sulit diinterpretasikan.

Kedua, ada risiko **kerapuhan model terhadap perubahan input**. Pedoman yang sama menyinggung bahwa kompleksitas model, arsitektur, dan ukuran parameter mempengaruhi robustness terhadap perubahan input saat sistem dipakai. Ini berarti model bisa tampil baik pada data tertentu, tetapi hasilnya menurun ketika pola input berubah.

Ketiga, ada risiko **ekstraksi informasi sensitif dari model**. Pada dokumen NIST AI 600-1, referensi yang dihimpun menunjukkan perhatian besar terhadap memorization, extraction of training data, dan theft of model behavior pada model bahasa modern. Walaupun dokumen tersebut pada potongan yang tersedia banyak memuat referensi, keberadaan referensi semacam ini sendiri menunjukkan bahwa model dapat menjadi target eksploitasi, bukan hanya alat yang pasif.

Keempat, ada risiko **perubahan perilaku model akibat pembaruan**. ISO/IEC 22989 menjelaskan bahwa AI system release management berbeda dari software biasa karena perubahan tidak hanya terjadi pada code, tetapi juga pada model dan bahkan training data diffs. Artinya, pembaruan model dapat mengubah perilaku sistem secara signifikan, sehingga perlu dikelola, diuji, dan didokumentasikan dengan hati-hati.

Kelima, ada risiko **ketergantungan pada model yang kurang dapat dijelaskan**. Bila organisasi tidak cukup memahami bagaimana model dipilih dan apa keterbatasannya, maka pengawasan menjadi lemah. Ini penting terutama bila model dipakai pada konteks yang mempengaruhi layanan, pengelompokan kasus, atau orientasi kebijakan.

Karena itu, risiko model bukan hanya risiko bahwa model “salah”. Risiko model juga berarti bahwa **organisasi mungkin tidak cukup memahami, tidak cukup mengendalikan, atau tidak cukup siap menghadapi perubahan perilaku model**.

5.3.3 Risiko pada output

Risiko ketiga berada pada **output** atau keluaran AI. Dalam banyak kasus, inilah bagian yang paling langsung dilihat pengguna, dan justru karena itu sering diberi kepercayaan berlebihan. Padahal, output AI sendiri adalah sumber risiko yang sangat penting.

Pertama, output bisa **meyakinkan tetapi keliru**. Ini berlaku pada ringkasan, jawaban, klasifikasi, rekomendasi, maupun draft awal. **ISO/IEC 22989** menegaskan bahwa keluaran AI bersifat *error-prone* dan pengguna perlu memahami implikasi akuntabilitas dari hasil yang salah.

Kedua, output bisa **melampaui ruang lingkup**. Misalnya, AI yang seharusnya hanya membantu orientasi awal justru menghasilkan jawaban yang terdengar seperti keputusan final. Atau AI yang seharusnya hanya merangkum malah menambahkan interpretasi baru yang tidak ada di sumber.

Ketiga, output dapat **membocorkan informasi**. Dalam solusi berbasis dokumen, retrieval, atau generative AI, keluaran dapat tanpa sengaja menampilkan informasi yang seharusnya tidak muncul kepada pengguna tertentu bila kontrol akses dan desain output tidak cukup baik.

Keempat, output bisa **mempengaruhi tindakan manusia secara berlebihan**. Ini penting karena dalam organisasi publik, hasil AI sering dipakai untuk:

- memulai tindak lanjut,
- memilih prioritas,
- menyiapkan briefing,
- atau menyusun tanggapan awal.

Jika manusia terlalu percaya pada output, maka kesalahan output dapat dengan mudah merembet menjadi kesalahan proses.

Kelima, output dapat **merusak integritas informasi**. Misalnya, ringkasan AI memadatkan dokumen tetapi menghilangkan syarat penting, nuansa kebijakan, atau konteks hukum yang justru krusial. Dalam situasi seperti ini, output tidak harus sepenuhnya salah untuk menjadi berbahaya; cukup dengan kehilangan bagian penting, ia sudah bisa menyesatkan.

Karena itu, organisasi perlu melihat output AI bukan hanya sebagai manfaat, tetapi juga sebagai sesuatu yang perlu:

- dibatasi,
- diperiksa,
- diberi label dengan jelas,
- dan diposisikan sesuai konteks penggunaannya.

5.3.4 Risiko pada prompt

Risiko keempat berada pada **prompt**. Ini adalah salah satu pembeda penting antara AI modern dan banyak sistem biasa. Dalam sistem generative AI dan agentic AI, prompt bukan hanya masukan data, tetapi juga instruksi yang membentuk perilaku sistem.

Risiko pertama pada prompt adalah **prompt yang buruk atau terlalu umum**. Ini mungkin terdengar lebih seperti masalah kualitas penggunaan daripada keamanan, tetapi dampaknya bisa nyata. Prompt yang tidak memberi konteks, batasan, atau format yang cukup dapat membuat AI menghasilkan output yang terlalu luas, terlalu bebas, atau tidak sesuai kebutuhan organisasi.

Risiko kedua adalah **prompt injection**. Ini adalah risiko ketika instruksi yang tidak semestinya disisipkan untuk memanipulasi perilaku sistem. Referensi yang dihimpun dalam NIST AI 600-1 juga memasukkan prompt injection sebagai isu penting dalam GAI, dan pedoman keamanan AI menekankan pentingnya sanitasi input serta guardrails terhadap interaksi model. Dalam solusi yang terhubung ke dokumen, plugin, atau tool eksternal, risiko ini menjadi lebih nyata karena sistem bisa diarahkan untuk mengabaikan batas yang semula ditetapkan.

Risiko ketiga adalah **system prompt atau meta-prompt yang tidak terdokumentasi dengan baik**. Pedoman keamanan AI menekankan perlunya mendokumentasikan system prompts, guardrails, scope, limitations, dan failure modes. Ini penting karena perubahan kecil pada prompt sistem dapat menghasilkan perubahan besar pada perilaku keluaran.

Risiko keempat adalah **kebocoran prompt atau kebocoran logika penggunaan**. Bila prompt internal yang sensitif, aturan respons, atau instruksi sistem terekspos, maka pihak yang tidak berwenang bisa lebih mudah mengeksploitasi atau memanipulasi perilaku AI.

Risiko kelima adalah **inkonsistensi penggunaan prompt**. Bila tiap pengguna memakai prompt yang sangat berbeda untuk tugas yang sama tanpa pedoman yang cukup, maka kualitas hasil menjadi tidak stabil dan pengendalian organisasi terhadap AI menjadi lemah.

Dengan demikian, prompt perlu diperlakukan sebagai aset organisasi yang perlu:

- dirancang,
- dibatasi,
- didokumentasikan,
- dan pada konteks tertentu diperlakukan dengan kontrol perubahan yang cukup disiplin.

5.3.5 Risiko pada vendor dan pihak ketiga

Risiko kelima berada pada **vendor** atau pihak ketiga. Dalam banyak solusi AI modern, organisasi memanfaatkan model, API, platform, plugin, atau layanan yang disediakan pihak luar. Ini dapat mempercepat implementasi, tetapi juga menambah titik risiko.

Pertama, ada risiko **ketergantungan**. Organisasi mungkin sangat bergantung pada vendor untuk model, pembaruan, ketersediaan layanan, atau fitur pengamanan tertentu. Jika vendor berubah kebijakan, mengubah model, atau mengalami insiden, organisasi bisa terdampak langsung.

Kedua, ada risiko **kurangnya visibilitas**. Bila komponen AI inti berasal dari pihak ketiga, organisasi mungkin tidak cukup memahami bagaimana model bekerja, bagaimana data diproses, atau bagaimana perubahan dilakukan dari waktu ke waktu.

Ketiga, ada risiko **privasi dan keamanan data**. NIST AI 600-1 dan pedoman keamanan AI sama-sama menekankan pentingnya dokumentasi, logging, dan review terhadap third-party inputs, plugins, serta pihak ketiga yang memiliki akses ke konten organisasi. NIST AI 600-1 juga menegaskan pentingnya panduan pelaporan insiden, change management records, version history, dan metadata dalam konteks GAI serta plugin pihak ketiga.

Keempat, ada risiko **supply chain**. Pedoman keamanan AI secara jelas menyebut *secure your supply chain* dan menekankan perlunya mempertimbangkan provenance serta supply chains dari model, training data, dan associated tools. Ini berarti bahwa vendor tidak hanya dipandang sebagai penyedia layanan, tetapi sebagai bagian dari rantai risiko AI.

Kelima, ada risiko **perubahan yang tidak terkendali**. Jika vendor memperbarui model atau layanan tanpa organisasi cukup siap, maka perilaku sistem bisa berubah di lapangan tanpa penyesuaian proses, dokumen, atau pengawasan yang memadai.

Karena itu, tata kelola vendor pada AI tidak cukup hanya berupa kontrak pengadaan biasa. Organisasi perlu menilai:

- data apa yang dibagikan,
- perubahan apa yang bisa terjadi,
- bagaimana logging dan audit dilakukan,
- siapa yang bertanggung jawab jika terjadi insiden,
- dan bagaimana fallback diterapkan bila layanan pihak ketiga tidak tersedia.

Penting untuk dipahami bahwa risiko pada data, model, output, prompt, dan vendor **tidak berdiri sendiri**. Mereka saling berkaitan. Misalnya:

- data yang buruk dapat membuat model memberi output yang buruk;
- prompt yang tidak aman dapat membuat output membocorkan data;
- vendor yang mengubah model dapat mengubah perilaku output;
- model yang rapuh dapat lebih mudah dipicu oleh prompt bermasalah;
- dan output yang menyesatkan dapat memicu tindakan manusia yang salah.

Cara pandang terhubung ini penting agar organisasi tidak mengelola risiko secara terpisah-pisah. Pengamanan pada satu titik perlu dipahami dalam hubungannya dengan titik lain.

Bagi organisasi, pembahasan ini dapat diterjemahkan menjadi beberapa pertanyaan praktis:

- Data apa yang paling menentukan perilaku sistem ini?
- Apakah model yang dipakai sesuai dengan use case dan cukup dapat diawasi?
- Apa bentuk output yang paling berisiko menyesatkan pengguna?
- Prompt atau instruksi apa yang perlu distandarkan atau dibatasi?
- Vendor atau layanan pihak ketiga apa saja yang memiliki pengaruh besar terhadap sistem?
- Jika salah satu unsur berubah, bagian mana dari workflow yang paling terdampak?

Pertanyaan-pertanyaan ini membantu organisasi membangun risk register yang lebih konkret dan tidak terlalu abstrak.

Risiko pada solusi AI dapat dipetakan secara lebih jelas pada lima titik penting: data, model, output, prompt, dan vendor. Data dapat salah, bias, tercemar, atau sensitif. Model dapat tidak cocok, rapuh, sulit dijelaskan, atau berubah perilakunya. Output dapat meyakinkan tetapi keliru, membocorkan informasi, atau menyesatkan tindakan manusia. Prompt dapat buruk, dimanipulasi, tidak terdokumentasi, atau tidak konsisten. Vendor dapat menambah ketergantungan, mengurangi visibilitas, memperluas risiko data, dan menjadi bagian dari supply chain risk.

5.4 Privasi dan perlindungan data

Di antara berbagai isu keamanan, etika, dan tata kelola AI, salah satu yang paling penting dalam sektor publik adalah **privasi dan perlindungan data**. Hal ini wajar, karena banyak use case AI bekerja di atas data yang berasal dari pelayanan, administrasi, komunikasi publik, dokumen internal, log sistem, atau pengetahuan kelembagaan. Sebagian data tersebut dapat mengandung informasi pribadi, data sensitif, informasi strategis organisasi, atau kombinasi semuanya sekaligus. Karena itu, ketika organisasi menggunakan AI, pertanyaan tentang data tidak cukup berhenti pada “apakah datanya tersedia” atau “apakah datanya cukup untuk model”. Organisasi juga harus bertanya: **data apa yang sedang diproses, siapa yang dapat mengaksesnya, bagaimana data itu dilindungi, dan apakah penggunaannya tetap menghormati privasi serta sesuai dengan tata kelola yang benar.**

Dalam konteks kebijakan nasional, **Peta Jalan Kecerdasan Artifisial Nasional Tahun 2026–2029** menempatkan **perlindungan data pribadi** sebagai salah satu prinsip pelaksanaan AI. Dokumen tersebut juga menempatkan keamanan, akuntabilitas, integritas, transparansi, dan keterlibatan manusia sebagai prinsip yang harus berjalan bersama. Sementara itu, **Konsep Pedoman Etika Kecerdasan Artifisial** menegaskan bahwa AI perlu dikembangkan, diselenggarakan, dan dimanfaatkan secara bertanggung jawab, transparan, menghormati privasi, menjamin keamanan, dan memungkinkan pengawasan manusia. Pedoman tersebut juga menekankan bahwa tata kelola AI perlu mencakup identifikasi dan mitigasi risiko serta langkah-langkah perlindungan yang memadai.

Dari sudut pandang keamanan sistem, **Guidelines for Secure AI System Development** memberi penekanan yang sangat kuat pada pengendalian data. Dokumen ini menyatakan bahwa organisasi perlu menerapkan kontrol yang tepat terhadap data yang dikirim ke luar kendali organisasi, melakukan pemeriksaan dan sanitasi input, membatasi data yang dapat diakses sistem AI, membatasi tindakan yang dapat dipicu AI, dan menerapkan prinsip *least privilege* agar komponen AI hanya mendapat akses minimum yang diperlukan untuk menjalankan fungsinya. Dokumen yang sama juga menekankan perlunya dokumentasi tentang sumber data, ruang lingkup, batas penggunaan, retensi, frekuensi peninjauan, serta keterbatasan dan failure modes dari penggunaan data dalam AI.

Dengan demikian, privasi dan perlindungan data dalam AI tidak dapat dipahami hanya sebagai kewajiban hukum atau administrasi. Ia merupakan bagian inti dari desain, integrasi, dan operasi sistem AI itu sendiri. Tanpa perlindungan data yang memadai, AI tidak hanya berisiko menimbulkan kebocoran atau pelanggaran privasi, tetapi juga berisiko merusak kepercayaan publik, memperlemah legitimasi organisasi, dan membuat manfaat yang diharapkan dari AI menjadi tidak sebanding dengan kerugian yang ditimbulkannya.

Privasi selalu penting dalam sistem digital, tetapi pada AI sensitivitasnya meningkat karena beberapa alasan.

- **Pertama**, AI sering bekerja dengan **volume data yang besar**. Untuk menjalankan fungsi seperti klasifikasi, ringkasan, retrieval, atau prediksi, sistem mungkin mengakses banyak dokumen atau banyak catatan sekaligus. Semakin besar volume dan keragaman data yang terlibat, semakin besar pula risiko bahwa informasi pribadi atau informasi sensitif ikut terproses dalam jumlah yang tidak kecil.
- **Kedua**, AI cenderung bekerja dengan **korelasi dan pola**. Ini berarti bahwa meskipun sebuah data tampak tidak terlalu sensitif secara mandiri, ketika dipadukan dengan data lain, ia bisa mengungkapkan informasi yang lebih dalam tentang individu, kelompok, atau perilaku tertentu. Dalam konteks organisasi publik, hal seperti ini bisa menyentuh identitas, layanan yang diterima seseorang, interaksi administratif, atau bahkan pola penggunaan sistem oleh pegawai.
- **Ketiga**, AI modern—terutama generative AI—dapat menghasilkan **output dalam bahasa alami**. Ini membuat kebocoran atau penyebutan informasi sensitif dapat muncul dalam bentuk yang lebih halus dan lebih mudah lolos dari perhatian pengguna dibanding sekadar tampilan mentah dari database.
- **Keempat**, dalam banyak use case, AI tidak hanya memproses data secara lokal. Ada situasi di mana organisasi memakai model pihak ketiga, API eksternal, plugin, atau layanan vendor. Dalam kondisi seperti ini, data dapat keluar dari lingkungan kendali langsung organisasi, sehingga risiko privasi dan perlindungan data meningkat bila kontrolnya tidak ketat.

Karena itu, penggunaan AI membuat isu privasi lebih dari sekadar pertanyaan “siapa yang boleh melihat data”. Ia juga mencakup:

- siapa yang boleh memproses data,
- dalam konteks apa data boleh dipakai,
- seberapa banyak data yang benar-benar perlu diberikan ke sistem,
- dan apakah AI dapat mengungkapkan kembali sesuatu yang seharusnya tidak muncul.

5.4.1 Privasi dalam Data

Salah satu kesalahan umum adalah menganggap bahwa privasi hanya relevan bila sistem memproses data yang secara eksplisit jelas merupakan data pribadi, seperti nama, NIK, alamat, atau nomor telepon. Dalam kenyataannya, isu privasi dalam AI jauh lebih luas.

Ada data yang tampak administratif atau operasional, tetapi bila dikombinasikan, bisa mengungkapkan informasi tentang:

- individu tertentu,
- pola interaksi seseorang dengan layanan,
- kondisi organisasi,
- atau perilaku pegawai dan pengguna.

Misalnya:

- dokumen pengaduan dapat memuat cerita yang mengidentifikasi seseorang secara tidak langsung;
- log interaksi sistem bisa menunjukkan pola aktivitas yang sensitif;
- kumpulan dokumen internal bisa mengandung korespondensi atau catatan yang seharusnya tidak diproses bebas;
- dan jawaban retrieval yang menggabungkan beberapa sumber bisa tanpa sengaja mengungkapkan informasi yang lebih banyak daripada yang seharusnya diperlukan.

Dengan kata lain, perlindungan privasi pada AI perlu mempertimbangkan bukan hanya data yang “jelas pribadi”, tetapi juga konteks, kombinasi data, dan potensi inferensi yang bisa dihasilkan sistem.

Salah satu prinsip paling penting dalam perlindungan data pada AI adalah **data minimization** dalam praktik, yaitu hanya menggunakan data yang benar-benar diperlukan untuk use case yang sedang dijalankan. Organisasi perlu membiasakan diri bertanya:

- data apa yang benar-benar dibutuhkan agar use case ini bekerja?
- apakah semua dokumen ini perlu dimasukkan ke sistem?
- apakah seluruh atribut dalam dataset perlu digunakan?
- apakah ada bagian yang bisa dihapus, dibatasi, atau disamarkan?

Prinsip ini penting karena banyak organisasi tergoda memasukkan sebanyak mungkin data ke AI dengan anggapan bahwa lebih banyak data selalu lebih baik. Pada kenyataannya, lebih banyak data juga berarti lebih banyak risiko. Bila manfaat dari data tambahan tidak jelas, maka memasukkannya justru bisa memperluas area paparan tanpa nilai yang sepadan.

Dalam sektor publik, pendekatan minimal seperti ini sangat penting. Organisasi sebaiknya tidak memberi AI akses ke data yang tidak relevan hanya karena data tersebut tersedia.

5.4.2 Kontrol Akses Data

Privasi dan perlindungan data dalam AI sangat bergantung pada **kontrol akses**. Ini berarti organisasi perlu menentukan dengan cukup ketat:

- siapa yang boleh mengakses data sumber;
- siapa yang boleh melihat output AI;
- siapa yang boleh menggunakan fungsi tertentu;
- dan sumber mana yang boleh dihubungkan ke sistem AI.

Pedoman keamanan AI menekankan pentingnya membatasi data yang dapat diakses sistem AI dan menerapkan prinsip *least privilege* agar sistem hanya memperoleh akses minimum yang dibutuhkan untuk fungsinya.

Dalam konteks organisasi, ini berarti misalnya:

- knowledge assistant internal tidak boleh menampilkan dokumen yang tidak berhak diakses pengguna;
- sistem ringkasan tidak boleh memproses repositori sensitif tanpa alasan yang jelas;
- plugin atau vendor pihak ketiga tidak boleh menerima data melebihi yang dibutuhkan;
- dan agentic AI tidak boleh diberi akses lintas aplikasi tanpa pembatasan yang jelas.

Kontrol akses ini harus diterapkan bukan hanya pada level aplikasi, tetapi juga pada level data, repositori, fungsi AI, dan bahkan hasil AI yang ditampilkan.

Dalam beberapa use case, organisasi dapat mengurangi risiko privasi dengan menerapkan langkah-langkah seperti:

- **de-identification**,
- **masking**,
- **redaction**,
- atau pembatasan isi dokumen sebelum diproses AI.

ISO/IEC 22989 menjelaskan bahwa data dalam AI dapat melalui proses seperti de-identification, filtering, normalization, dan data quality checking sebagai bagian dari persiapan dan pengelolaan data. Ini penting karena tidak semua AI harus memproses data dalam bentuk paling lengkap.

Sebagai contoh:

- untuk analisis tema pengaduan, organisasi mungkin tidak perlu memasukkan identitas penuh pelapor;
- untuk ringkasan laporan, bagian tertentu bisa disamarkan terlebih dahulu;
- untuk evaluasi pola, atribut identitas dapat dibatasi bila tidak relevan;
- dan untuk pengujian, data kasus uji bisa dibuat representatif tanpa harus menggunakan data produksi yang sepenuhnya utuh.

Pendekatan ini membantu organisasi mengurangi paparan yang tidak perlu sambil tetap mempertahankan nilai analitis dari data.

5.4.3 Risiko kebocoran data melalui output AI

Salah satu risiko yang khas pada AI adalah bahwa kebocoran data tidak selalu terjadi hanya pada level database atau file. Ia juga bisa terjadi melalui **output AI**. Ini penting terutama pada generative AI dan retrieval-based systems.

Misalnya:

- sistem bisa menyebut kembali detail yang seharusnya tidak ditampilkan;
- jawaban berbasis dokumen bisa menggabungkan potongan informasi sensitif ke dalam satu keluaran;
- ringkasan bisa memunculkan nama, konteks, atau keterangan yang sebenarnya tidak perlu;
- atau model dapat “mengingat” dan mengulang pola tertentu dari data yang pernah diproses.

Karena itu, perlindungan data dalam AI tidak cukup hanya memeriksa siapa yang boleh mengakses input. Organisasi juga perlu menilai:

- apa yang mungkin keluar sebagai output;
- bagaimana output dibatasi;
- apakah sumber yang terlalu sensitif boleh diringkas atau tidak;
- dan apakah pengguna diberi hasil dalam bentuk yang proporsional dengan hak akses dan kebutuhannya.

Dengan kata lain, **privasi dalam AI juga adalah masalah desain output**, bukan hanya masalah penyimpanan data.

Privasi menjadi lebih sensitif ketika organisasi memakai vendor, model pihak ketiga, atau layanan API eksternal. Dalam situasi seperti ini, organisasi tidak hanya perlu memikirkan bagaimana data diolah di dalam sistem sendiri, tetapi juga bagaimana data keluar ke pihak luar.

NIST AI 600-1 menekankan pentingnya inventarisasi pihak ketiga yang memiliki akses ke konten organisasi, evaluasi risiko pemasok, due diligence atas privasi dan keamanan, pencatatan perubahan layanan, serta rencana kontinjensi bila layanan pihak ketiga mengalami masalah.

Bagi organisasi pemerintah, ini berarti perlu ada pertanyaan yang lebih rinci seperti:

- data apa yang dikirim ke layanan eksternal;
- apakah data itu benar-benar perlu;
- apakah ada bagian yang bisa dihapus atau disamarkan lebih dulu;
- apa kebijakan retensi vendor terhadap data tersebut;
- apakah vendor memakai data untuk pelatihan lebih lanjut atau tidak;
- dan apa kontrol yang dimiliki organisasi bila ingin membatasi atau menghentikan pemrosesan.

Tanpa pertanyaan-pertanyaan ini, organisasi dapat secara tidak sadar memperluas risiko privasi jauh di luar batas yang mereka pahami.

Dalam sektor publik, pelanggaran atau kelemahan perlindungan data pada AI bukan hanya masalah kepatuhan atau keamanan teknis. Ia juga masalah **kepercayaan**. Masyarakat, pegawai, dan pemangku kepentingan perlu yakin bahwa penggunaan AI oleh instansi:

- tidak memanfaatkan data secara berlebihan;
- tidak memperlakukan data pribadi secara serampangan;
- dan tidak membuka risiko baru yang sebenarnya dapat dicegah.

Kepercayaan ini penting karena AI seringkali melibatkan bentuk pemrosesan yang tidak langsung tampak oleh pengguna. Jika organisasi tidak dapat menjelaskan bagaimana data dilindungi dan dibatasi, maka penerimaan terhadap AI juga akan melemah.

Pelajaran penting yang perlu ditekankan adalah bahwa privasi dan perlindungan data tidak boleh ditempelkan belakangan setelah sistem selesai dibuat. Keduanya perlu dibangun **sejak desain awal**. Ini berarti sejak tahap use case, organisasi sebaiknya sudah bertanya:

- data apa yang akan diproses;
- apakah ada data pribadi atau data sensitif;
- adakah alternatif data yang lebih aman;
- kontrol akses seperti apa yang dibutuhkan;
- apakah ada kebutuhan de-identification;
- dan bagaimana output akan dibatasi.

Pendekatan seperti ini membuat privasi menjadi bagian dari desain sistem, bukan sekadar pengecekan setelah masalah muncul.

5.4.4 Peran dokumentasi dalam perlindungan data

Karena AI sering melibatkan banyak komponen, dokumentasi juga menjadi sangat penting dalam isu privasi. Organisasi sebaiknya mendokumentasikan:

- sumber data;
- jenis data yang dipakai;
- tujuan penggunaan;
- batas penggunaan;
- retensi data;
- pihak yang memiliki akses;
- dan kontrol apa yang diterapkan.

Pedoman keamanan AI secara jelas mendukung dokumentasi semacam ini sebagai bagian dari transparansi dan akuntabilitas.

Dengan dokumentasi yang baik, organisasi akan lebih siap untuk:

- menjelaskan desain pengendalian;
- meninjau ulang penggunaan data;
- dan melakukan perbaikan bila ditemukan risiko baru.

Pada akhirnya, perlindungan data dalam AI juga memerlukan **human oversight**. Pengguna dan pengelola sistem perlu memahami:

- jenis data apa yang boleh dan tidak boleh dimasukkan ke AI;
- kapan hasil AI perlu diperiksa agar tidak mengungkap terlalu banyak;
- dan kapan proses harus dihentikan karena konteks datanya terlalu sensitif.

Tanpa pemahaman manusia yang cukup, bahkan kontrol teknis yang baik pun bisa dilewati dalam praktik. Karena itu, privasi dan perlindungan data perlu dibaca bersama pelatihan, kebijakan internal, dan pengawasan kerja sehari-hari.

Privasi dan perlindungan data adalah pondasi penting dalam penggunaan AI di sektor publik. AI tidak hanya memproses data, tetapi dapat memperluas cara data dibaca, digabungkan, dan ditampilkan. Karena itu, organisasi tidak cukup hanya memikirkan ketersediaan data, tetapi juga harus menilai kebutuhan data, kontrol akses, de-identification, pembatasan output, pengelolaan vendor, dokumentasi, dan pengawasan manusia.

5.5 Bias, fairness, dan integritas informasi

Setelah membahas privasi dan perlindungan data, pembahasan keamanan, etika, dan tata kelola AI perlu dilanjutkan ke tiga hal yang sangat penting dalam sektor publik, yaitu **bias**, **fairness**, dan **integritas informasi**. Ketiganya perlu dibahas bersama karena dalam praktik penggunaan AI, masalah tidak selalu muncul dalam bentuk kebocoran data atau kegagalan teknis yang mudah terlihat. Kadang sistem tampak berjalan dengan baik, hasilnya tampak rapi, dan proses terasa lebih cepat, tetapi di bawah permukaan terdapat persoalan lain: sistem bisa cenderung tidak adil, memperkuat ketimpangan yang sudah ada, atau menghasilkan informasi yang tampak meyakinkan tetapi mengganggu kualitas pemahaman organisasi.

Dalam konteks sektor publik, hal ini sangat penting. Pemerintah tidak hanya membutuhkan sistem yang cepat atau efisien, tetapi juga sistem yang:

- memperlakukan kasus secara wajar,
- tidak mendorong diskriminasi yang tidak semestinya,
- tidak mempersempit pandangan organisasi secara keliru,
- dan tidak merusak mutu informasi yang menjadi dasar kerja, layanan, atau analisis.

Karena itu, ketika organisasi menggunakan AI, pertanyaan yang perlu diajukan bukan hanya:

- apakah sistem ini bekerja, tetapi juga:
- apakah sistem ini bekerja secara cukup adil,
- apakah ada kelompok, wilayah, jenis dokumen, atau jenis kasus tertentu yang lebih sering dirugikan,
- dan apakah hasil AI membantu memperkuat pemahaman atau justru menciptakan gambaran yang menyesatkan.

Konsep Pedoman Etika Kecerdasan Artifisial secara tegas menyebut tantangan seperti **bias algoritma**, pelanggaran penggunaan data, dan disinformasi sebagai bagian dari tantangan etika AI yang perlu ditangani. Dokumen tersebut juga menempatkan keadilan, transparansi, keamanan, penghormatan terhadap privasi, serta pengawasan manusia sebagai bagian dari prinsip penggunaan AI yang bertanggung jawab. Sementara itu, **Peta Jalan Kecerdasan Artifisial Nasional Tahun 2026–2029** menempatkan **keadilan**, **integritas**, **inklusivitas**, dan **akuntabilitas** sebagai prinsip pelaksanaan AI.

Pada tingkat standar, **ISO/IEC 22989** juga menempatkan bias, keamanan digital, privasi, keselamatan, dan manajemen risiko sebagai unsur yang perlu diperhatikan di sepanjang siklus hidup AI. Standar ini bahkan menegaskan bahwa AI system life cycle memerlukan pendekatan manajemen risiko yang berkelanjutan dan mencakup bias serta keamanan sebagai bagian dari pertimbangan sistem. Dengan demikian, bias, fairness, dan integritas informasi bukan isu pinggiran. Ketiganya adalah bagian inti dari apakah AI layak digunakan dalam organisasi publik.

Dalam pembicaraan sehari-hari, istilah **bias** sering dipahami sebagai keberpihakan yang salah atau kecenderungan yang tidak netral. Dalam konteks AI, bias dapat dipahami sebagai kecenderungan sistem untuk menghasilkan perlakuan, keluaran, atau pola yang tidak seimbang, tidak proporsional, atau tidak semestinya terhadap kelompok, jenis kasus, kategori data, atau konteks tertentu.

Bias tidak selalu berarti sistem “sengaja diskriminatif”. Dalam banyak kasus, bias muncul karena:

- data yang dipakai tidak cukup mewakili kenyataan yang beragam;
- label atau kategori historis sudah membawa ketimpangan lama;
- desain sistem terlalu menyederhanakan variasi kasus;
- pengguna memakai AI pada konteks yang melampaui tujuan awal;
- atau evaluasi hanya dilakukan pada rata-rata hasil tanpa melihat variasi antar jenis kasus.

Sebagai contoh:

- sistem klasifikasi aduan bisa lebih sering salah pada jenis bahasa tertentu;
- retrieval dokumen bisa lebih sering menampilkan sumber dari kategori yang dominan dan mengabaikan sumber yang lebih relevan tetapi kurang sering muncul;
- ringkasan AI bisa lebih mudah menangkap isu yang eksplisit dan mengabaikan isu minoritas yang tertulis lebih halus;
- atau model penandaan prioritas bisa mereplikasi pola lama yang sebenarnya tidak selalu adil.

Dari sini terlihat bahwa bias dalam AI tidak selalu dramatis, tetapi bisa hadir sebagai **ketidakseimbangan sistematis** yang perlahan mempengaruhi kualitas layanan, analisis, atau keputusan awal organisasi.

5.5.1 Kemunculan Bias

Penting untuk dipahami bahwa bias tidak berasal dari satu tempat saja. Ia bisa muncul dari banyak titik dalam solusi AI.

- Pertama, bias bisa muncul dari **data**. Bila data historis tidak representatif, tidak lengkap, atau sudah memuat ketimpangan lama, maka AI bisa mewarisi pola tersebut.
- Kedua, bias bisa muncul dari **definisi kategori dan label**. Bila organisasi memakai kategori yang terlalu kasar, terlalu tumpang tindih, atau tidak sensitif terhadap variasi kasus, maka hasil AI juga akan mudah condong ke pola yang tidak adil.
- Ketiga, bias bisa muncul dari **desain model atau pipeline**. Cara memilih fitur, cara menentukan threshold, atau cara menyusun retrieval corpus dapat mempengaruhi keluaran.
- Keempat, bias bisa muncul dari **prompt dan instruksi**. Pada generative AI, prompt yang terlalu sempit atau terlalu mengarahkan bisa membuat sistem mengabaikan dimensi lain yang seharusnya ikut dipertimbangkan.
- Kelima, bias bisa muncul dari **cara pengguna menafsirkan output AI**. Bahkan bila sistem tidak sepenuhnya bias, pengguna bisa memberi bobot berlebihan pada hasil tertentu dan mengabaikan keterbatasannya.

Karena itu, organisasi sebaiknya tidak melihat bias hanya sebagai masalah model. Bias perlu dibaca sebagai persoalan yang dapat muncul di seluruh rantai penggunaan AI.

5.5.2 Fairness dalam AI

Istilah **fairness** atau keadilan dalam AI perlu dipahami secara hati-hati. Fairness bukan berarti semua orang, semua dokumen, atau semua kasus harus selalu diperlakukan identik tanpa melihat konteks. Dalam sektor publik, fairness lebih dekat dengan gagasan bahwa sistem tidak boleh menghasilkan perlakuan yang tidak semestinya, tidak proporsional, atau tidak dapat dibenarkan terhadap individu, kelompok, jenis kasus, atau konteks tertentu.

Ini berarti fairness menuntut organisasi untuk bertanya:

- apakah sistem bekerja cukup baik di berbagai jenis input yang relevan;
- apakah ada kelompok atau konteks tertentu yang secara konsisten dirugikan;
- apakah hasil sistem menambah hambatan bagi pihak yang sudah lebih rentan;
- dan apakah hasil AI dapat dijelaskan serta dikoreksi bila menimbulkan ketidakadilan.

Dalam praktik, fairness sering lebih tepat dipahami sebagai **kewaspadaan terhadap ketidakwajaran pola hasil**, bukan sekadar pencarian angka tunggal. Pada banyak use case sektor publik, hal yang penting adalah memastikan bahwa:

- AI tidak memperbesar bias yang sudah ada;
- AI tidak mempermudah perlakuan yang tidak konsisten;
- dan AI tidak mendorong organisasi mengambil jalan pintas yang mengorbankan keadilan.

Mengapa fairness begitu penting di sektor publik? Karena organisasi publik bekerja dalam ruang yang menyangkut layanan, hak, akses, dan kepercayaan masyarakat. Bila AI digunakan pada proses yang berkaitan dengan:

- pengelompokan aduan,
 - penyiapan informasi layanan,
 - pencarian dasar regulasi,
 - penandaan prioritas pemeriksaan,
 - atau analisis isu kebijakan,
- maka ketidakadilan dalam hasil AI bisa berdampak lebih luas daripada sekadar kesalahan teknis.

Misalnya:

- bila aduan tertentu lebih sering salah dikelompokkan, maka tindak lanjut bisa terlambat;
- bila pencarian dokumen internal lebih sering menampilkan dasar yang dominan dan mengabaikan pengecualian penting, maka interpretasi kebijakan bisa timpang;
- bila ringkasan masukan publik hanya menangkap tema mayoritas yang paling eksplisit, maka perspektif minoritas bisa hilang dari perhatian;
- bila deteksi anomali terlalu sering menandai pola tertentu tanpa konteks, tim pengawasan bisa teralihkan dari risiko yang sebenarnya lebih penting.

Karena itu, fairness dalam AI sektor publik bukan hanya isu statistik. Ia berkaitan dengan kualitas tata kelola dan legitimasi organisasi.

5.5.3 Integritas informasi

Selain bias dan fairness, satu isu yang tidak kalah penting adalah **integritas informasi**. Ini adalah kemampuan organisasi untuk menjaga agar informasi yang dipakai, dihasilkan, dirangkum, atau disebarkan melalui AI tetap:

- cukup akurat,
- cukup utuh,
- tidak menyesatkan,
- dan tidak kehilangan konteks penting.

Dalam konteks AI, integritas informasi penting karena banyak use case justru berkaitan langsung dengan pengolahan informasi, misalnya:

- ringkasan dokumen,
- tanya jawab berbasis dokumen,
- drafting awal,
- analisis tema,
- atau retrieval pengetahuan internal.

Pada semua use case itu, AI tidak sekadar menghitung angka. Ia menyentuh cara organisasi **membaca, memahami, dan menyampaikan informasi**. Bila integritas informasi terganggu, maka AI bisa membuat organisasi bergerak lebih cepat, tetapi ke arah yang salah.

Contohnya:

- ringkasan yang terlalu pendek dapat menghilangkan syarat penting;
- jawaban berbasis dokumen bisa terdengar pasti padahal sumbernya ambigu;
- draft awal bisa tampak rapi tetapi menambahkan klaim yang tidak ada;
- analisis tema bisa terlalu menyederhanakan persoalan yang sebenarnya kompleks.

Dengan kata lain, integritas informasi mengingatkan bahwa keluaran AI tidak cukup dinilai dari kerapian bahasanya. Yang lebih penting adalah apakah keluaran itu menjaga hubungan yang sehat dengan fakta, sumber, konteks, dan makna yang seharusnya dipertahankan.

Pada generative AI, isu integritas informasi sangat erat dengan risiko **confabulation** atau keluaran yang tampak masuk akal tetapi tidak didasarkan pada sumber yang benar. Walaupun istilah ini akan dibahas lebih khusus pada subbagian risiko generative AI, di sini penting dicatat bahwa confabulation bukan hanya masalah “AI salah”. Ia adalah masalah integritas informasi karena sistem dapat:

- mengisi celah informasi dengan hal yang tidak ada,
- menyusun kalimat yang terdengar otoritatif,
- dan membuat pengguna merasa seolah-olah hasil tersebut dapat dipercaya.

Dalam organisasi publik, keluaran semacam itu sangat berbahaya bila tidak diperiksa. Karena itu, menjaga integritas informasi berarti juga menjaga agar AI:

- terhubung ke sumber yang benar,
- dibatasi ruang geraknya,
- dan diposisikan sebagai alat bantu yang tetap perlu diverifikasi.

Bias, fairness, dan integritas informasi sebenarnya saling berhubungan. Sebuah sistem yang tidak adil sering juga merusak integritas informasi, karena ia menyajikan gambaran yang tidak lengkap atau tidak proporsional tentang kenyataan. Sebaliknya, keluaran yang buruk integritas informasinya dapat mendorong perlakuan yang tidak adil, karena keputusan atau tindak lanjut diambil di atas dasar pemahaman yang salah.

Sebagai contoh:

- jika ringkasan masukan publik lebih banyak menangkap suara kelompok dominan, itu bukan hanya masalah bias, tetapi juga masalah integritas informasi;
- jika sistem knowledge assistant hanya sering menampilkan sebagian dokumen yang paling populer tetapi mengabaikan sumber yang lebih relevan, itu bukan hanya masalah retrieval quality, tetapi juga masalah fairness terhadap informasi yang tersedia;
- jika AI drafting membuat narasi yang terlalu yakin tanpa dasar, maka integritas informasi turun dan keputusan manusia juga bisa ikut terpengaruh.

Karena itu, organisasi sebaiknya tidak mengelola ketiga isu ini secara terpisah-pisah.

Agar lebih praktis, pembaca dapat melihat beberapa kemungkinan bentuk bias menurut jenis use case.

- Pada **klasifikasi**, bias bisa muncul bila kategori tertentu lebih sering diprediksi dengan baik daripada kategori lain, atau bila sistem kurang baik menangani variasi bahasa dan konteks tertentu.
- Pada **retrieval**, bias bisa muncul bila dokumen tertentu lebih mudah muncul hanya karena struktur atau frekuensinya, bukan karena relevansinya yang sebenarnya lebih tinggi.
- Pada **ringkasan**, bias bisa muncul bila AI cenderung menonjolkan isu yang paling eksplisit dan mengabaikan isu yang lebih halus atau minoritas.
- Pada **drafting**, bias bisa muncul bila gaya bahasa atau susunan argumen yang dihasilkan cenderung memproduksi perspektif tertentu dan mengabaikan kehati-hatian yang dibutuhkan.
- Pada **deteksi anomali**, bias bisa muncul bila sistem lebih sering menandai pola dari konteks tertentu tanpa dasar yang proporsional.

Melihat bias berdasarkan use case seperti ini membantu organisasi memikirkan langkah mitigasi yang lebih konkret.

Untuk menjaga fairness dan integritas informasi, organisasi dapat melakukan beberapa langkah dasar.

- Pertama, **memeriksa data dan sumber**. Organisasi perlu menilai apakah data cukup representatif, cukup mutakhir, dan cukup layak untuk konteks penggunaan yang direncanakan.
- Kedua, **menguji sistem pada variasi kasus**. Jangan hanya menguji hasil rata-rata. Lihat juga apakah ada jenis dokumen, jenis pertanyaan, kategori kasus, atau kelompok konteks tertentu yang hasilnya lebih lemah.
- Ketiga, **mewajibkan sumber atau jejak rujukan** pada use case tertentu. Ini terutama penting pada retrieval, tanya jawab berbasis dokumen, dan ringkasan berbasis sumber.
- Keempat, **menjaga human oversight**. Pengguna perlu tetap memeriksa apakah hasil AI masuk akal, adil, dan cukup setia pada sumber.
- Kelima, **mendokumentasikan keterbatasan sistem**. Pedoman keamanan AI menekankan pentingnya dokumentasi tentang keterbatasan, ruang lingkup, dan failure modes. Ini penting agar organisasi tidak menganggap hasil AI berlaku sama baik untuk semua konteks.
- Keenam, **menangkap koreksi dan umpan balik**. Bila pengguna sering memperbaiki hasil tertentu, itu adalah sinyal bahwa sistem punya kelemahan yang perlu dievaluasi.
- Ketujuh, **menghindari penggunaan AI sebagai penentu final** pada konteks yang fairness dan integritas informasinya sangat penting, kecuali organisasi benar-benar memiliki kontrol, pengujian, dan tata kelola yang memadai.

5.5.4 Fairness dan integritas informasi

Seperti isu risiko AI lain, fairness dan integritas informasi tidak dapat dipastikan sekali di awal lalu dianggap selesai. Perubahan pada:

- data,
 - model,
 - prompt,
 - vendor,
 - repositori dokumen,
 - atau perilaku pengguna
- dapat mengubah pola hasil AI. Karena itu, organisasi perlu memantau:
- koreksi pengguna,
 - pola kesalahan,
 - keluhan,
 - variasi performa antar jenis kasus,
 - dan perubahan kualitas informasi yang dihasilkan dari waktu ke waktu.

Ini sejalan dengan pendekatan **continuous validation** dan **re-evaluation** pada siklus hidup AI yang dijelaskan dalam ISO/IEC 22989.

Bias, fairness, dan integritas informasi adalah tiga isu yang sangat penting dalam penggunaan AI di sektor publik. Bias dapat membuat sistem menghasilkan pola yang tidak seimbang atau tidak semestinya. Fairness menuntut agar organisasi waspada terhadap hasil yang merugikan konteks atau kelompok tertentu secara tidak proporsional. Integritas informasi menuntut agar AI tidak merusak mutu informasi yang menjadi dasar kerja, layanan, dan analisis organisasi.

5.6 Risiko khusus generative AI dan agentic AI

Pada subbagian sebelumnya telah dibahas bias, fairness, dan integritas informasi sebagai isu penting dalam penggunaan AI di sektor publik. Namun, perkembangan AI dalam beberapa tahun terakhir membawa bentuk risiko yang semakin khas, terutama pada **generative AI** dan **agentic AI**. Risiko-risiko ini perlu dibahas secara khusus karena keduanya memiliki sifat kerja yang berbeda dari sistem berbasis aturan biasa maupun model analitik yang lebih terbatas.

Generative AI mampu menghasilkan keluaran baru dalam bentuk bahasa alami, ringkasan, draft, penjelasan, tabel, atau bentuk representasi lain yang tampak rapi dan meyakinkan. Sementara itu, agentic AI tidak hanya menghasilkan satu keluaran, tetapi dapat membantu menjalankan beberapa langkah kerja secara berurutan, misalnya mencari informasi, memilih sumber, merangkum, menyusun draft, lalu mengarahkan langkah berikutnya dalam suatu workflow. Kedua bentuk ini sangat bermanfaat, tetapi justru karena keduanya tampak fleksibel, cepat, dan mampu “membantu berpikir”, maka risikonya juga perlu dibaca dengan jauh lebih hati-hati.

Dalam organisasi pemerintah, generative AI dan agentic AI dapat dipakai untuk banyak hal: penyusunan bahan briefing, ringkasan dokumen, tanya jawab berbasis dokumen, drafting awal tanggapan, hingga workflow multi-langkah yang melibatkan beberapa sumber dan tool. Namun, penggunaan yang semakin luas ini juga membuka sejumlah risiko yang lebih khas, seperti:

- **confabulation**;
- **prompt injection**;
- **data leakage**;
- **over-automation**;
- dan **tindakan agent yang tidak terkendali**.

Kelima risiko ini bukan sekadar persoalan teknis. Masing-masing dapat berdampak pada kualitas kerja, keamanan informasi, keadilan proses, dan akuntabilitas organisasi. Karena itu, organisasi perlu memahaminya bukan hanya agar bisa “waspada”, tetapi agar bisa merancang kontrol, guardrail, dan pengawasan manusia yang lebih tepat.

Sistem berbasis aturan tradisional biasanya bekerja dalam ruang yang lebih sempit. Logika kerjanya relatif eksplisit, keluaran yang mungkin juga lebih terbatas, dan titik-titik perilakunya lebih mudah diprediksi. Generative AI berbeda. Ia menghasilkan keluaran dalam bahasa alami yang terasa luwes, kadang terlihat sangat masuk akal, dan sering membuat pengguna merasa sedang berhadapan dengan sistem yang “mengerti” konteks. Agentic AI bergerak lebih jauh lagi, karena tidak hanya menghasilkan satu jawaban, tetapi dapat menautkan beberapa langkah kerja.

Karena itu, risiko pada generative AI dan agentic AI bukan hanya soal apakah sistem memberi output yang salah. Resikonya juga menyangkut:

- bagaimana pengguna menafsirkan keluaran yang tampak meyakinkan;
- bagaimana input atau instruksi mempengaruhi perilaku sistem;
- bagaimana sistem dapat mengekspos informasi yang tidak semestinya;
- bagaimana manusia bisa terlalu cepat menyerahkan pekerjaan ke AI;
- dan bagaimana AI yang diberi akses ke tool atau workflow dapat melangkah lebih jauh dari yang dirancang.

Dengan kata lain, risiko pada generative AI dan agentic AI sangat berkaitan dengan **bahasa, konteks, kepercayaan, dan tindakan**. Inilah yang membuat keduanya perlu diperlakukan secara lebih hati-hati.

5.6.1 Confabulation

Risiko pertama yang sangat penting adalah **confabulation**. Dalam konteks generative AI, confabulation dapat dipahami sebagai situasi ketika sistem menghasilkan keluaran yang tampak masuk akal, rapi, dan meyakinkan, tetapi sebenarnya:

- tidak cukup didukung oleh sumber yang benar;
- menambahkan detail yang tidak ada;
- menyambungkan informasi secara tidak sah;
- atau menyajikan sesuatu seolah-olah pasti, padahal dasar faktualnya lemah atau tidak tersedia.

Confabulation berbeda dari kesalahan teknis biasa. Ia lebih berbahaya karena tampilannya sering tidak memberi sinyal yang jelas bahwa sistem sedang keliru. Justru, keluaran AI bisa terdengar sangat meyakinkan. Dalam organisasi publik, ini menjadi masalah yang serius karena keluaran seperti itu bisa dipakai untuk:

- briefing pimpinan;
- orientasi awal terhadap kebijakan;
- penyusunan tanggapan;
- penjelasan dokumen;
- atau dukungan layanan internal.

Sebagai contoh, knowledge assistant internal dapat menjawab pertanyaan dengan bahasa yang sangat meyakinkan, tetapi ternyata menyandarkan jawaban pada dokumen yang tidak tepat. Sistem ringkasan dapat menambahkan “kesimpulan” yang terdengar logis, padahal tidak pernah tertulis dalam dokumen sumber. Draft jawaban dapat tampak formal dan rapi, tetapi mengandung penjelasan yang tidak sesuai dengan kebijakan aktual.

Bahaya confabulation di sektor publik terletak pada dua hal. Pertama, organisasi bisa bergerak berdasarkan informasi yang salah. Kedua, manusia bisa mengurangi tingkat kehati-hatiannya karena keluaran AI terlihat lebih rapi daripada catatan mentah atau bahan awal.

Karena itu, generative AI tidak boleh diposisikan sebagai sumber kebenaran final. Keluaran AI perlu diperlakukan sebagai:

- bahan awal;
- bantuan orientasi;
- draft sementara;
- atau sintesis awal yang tetap harus diverifikasi.

Salah satu cara paling penting untuk mengurangi confabulation adalah menghubungkan generative AI dengan **sumber resmi** yang jelas, membatasi ruang lingkup tugasnya, meminta sumber rujukan bila relevan, dan menjaga agar manusia tetap memeriksa hasil pada titik-titik penting. Dalam konteks organisasi, confabulation juga harus dipahami sebagai risiko yang semakin besar ketika:

- pertanyaan terlalu ambigu;
- dokumen sumber tidak cukup baik;
- prompt terlalu umum;
- atau AI diminta melampaui informasi yang sebenarnya tersedia.

5.6.2 Prompt injection

Risiko kedua adalah **prompt injection**. Ini adalah salah satu risiko yang paling khas pada generative AI dan terutama agentic AI. Secara sederhana, prompt injection terjadi ketika sistem dipengaruhi oleh instruksi yang tidak semestinya, sehingga ia menyimpang dari tujuan, batas, atau aturan yang seharusnya dijaga.

Dalam sistem generative AI modern, input bukan hanya data biasa. Input juga dapat berfungsi sebagai instruksi. Karena itu, bila ada instruksi tersembunyi, manipulatif, atau menyesatkan yang masuk ke sistem—baik melalui prompt pengguna maupun melalui dokumen atau sumber yang dibaca sistem—maka sistem dapat terdorong untuk:

- mengabaikan aturan awal;
- mengubah fokus tugas;
- menampilkan informasi yang tidak seharusnya;
- atau menjalankan langkah yang tidak diinginkan.

Prompt injection dapat muncul dalam beberapa bentuk. Misalnya, pengguna secara langsung mencoba menulis instruksi untuk “mengabaikan aturan sebelumnya”. Pada sistem yang terhubung ke dokumen atau retrieval, prompt injection juga bisa muncul melalui teks di dalam dokumen yang dibaca AI, jika sistem tidak cukup membedakan antara **isi dokumen** dan **instruksi kepada model**. Pada agentic AI, risikonya bisa lebih besar karena satu instruksi yang berhasil mempengaruhi sistem dapat mendorong agent menjalankan langkah lanjutan yang tidak semestinya.

Dalam konteks organisasi pemerintah, prompt injection berbahaya karena dapat:

- melemahkan guardrail;
- membocorkan informasi;
- menyesatkan pencarian dan sintesis dokumen;
- atau mengganggu workflow yang sudah dirancang agar tetap berada dalam batas tertentu.

Karena itu, organisasi perlu memahami bahwa keamanan input pada generative AI bukan hanya soal memeriksa format data. Ia juga soal memeriksa **peran instruksi**. Sistem perlu dirancang agar:

- membedakan secara tegas antara instruksi sistem dan konten dokumen;
- membatasi tindakan agent berdasarkan aturan yang tidak mudah digeser oleh prompt biasa;
- menolak atau membatasi input yang berusaha menabrak kebijakan;
- dan, pada use case sensitif, tidak secara otomatis menjalankan langkah lebih jauh tanpa titik persetujuan manusia.

Prompt injection menjadi salah satu alasan mengapa agentic AI tidak boleh diberi ruang gerak terlalu luas tanpa pembatasan yang jelas.

5.6.3 Data leakage

Risiko ketiga adalah **data leakage** atau kebocoran data. Dalam konteks AI, kebocoran data tidak selalu terjadi dalam bentuk klasik seperti database yang diretas. Pada generative AI dan agentic AI, kebocoran data bisa terjadi melalui beberapa jalur yang lebih halus.

Pertama, kebocoran dapat terjadi ketika pengguna memasukkan data sensitif ke sistem AI tanpa kontrol yang memadai. Jika organisasi memakai layanan pihak ketiga atau model eksternal, maka data yang dimasukkan bisa berpindah ke luar lingkungan kendali langsung organisasi.

Kedua, kebocoran dapat terjadi melalui **output AI**. Sistem bisa tanpa sengaja menampilkan kembali informasi sensitif, menggabungkan beberapa sumber menjadi jawaban yang terlalu terbuka, atau menghasilkan ringkasan yang memuat detail yang seharusnya tidak keluar.

Ketiga, pada agentic AI, kebocoran dapat terjadi ketika agent diberi akses ke terlalu banyak sumber atau tool, lalu menggunakan akses tersebut di luar konteks yang sebenarnya diperlukan. Dalam situasi seperti ini, masalahnya bukan hanya sistem “bocor”, tetapi bahwa desain kewenangan agent terlalu longgar.

Keempat, data leakage juga bisa terjadi bila:

- log interaksi tidak dikelola dengan benar;
- prompt yang memuat informasi sensitif tersimpan tanpa perlindungan yang cukup;

- atau integrasi dengan plugin dan vendor tidak dibatasi secara disiplin.

Dalam sektor publik, kebocoran data melalui AI sangat serius karena dapat menyangkut:

- data pribadi warga;
- dokumen internal;
- korespondensi resmi;
- informasi kebijakan yang belum final;
- atau catatan organisasi yang memiliki nilai strategis.

Karena itu, penggunaan generative AI dan agentic AI harus selalu dibaca dengan prinsip:

- hanya gunakan data yang benar-benar diperlukan;
- batasi apa yang boleh dimasukkan ke sistem;
- batasi apa yang boleh diakses model atau agent;
- batasi apa yang boleh ditampilkan sebagai output;
- dan tentukan dengan jelas kapan layanan pihak ketiga boleh dipakai dan dengan syarat apa.

Dengan kata lain, data leakage pada AI adalah persoalan desain sistem, tata kelola data, hak akses, dan disiplin operasional sekaligus.

5.6.4 Over-automation

Risiko keempat adalah **over-automation**, yaitu kecenderungan menyerahkan terlalu banyak tugas, penilaian, atau langkah kerja kepada AI hanya karena AI tampak mampu melakukannya. Risiko ini sangat penting dibahas karena justru pada generative AI dan agentic AI, godaan untuk melakukan over-automation sangat besar.

Generative AI membuat pekerjaan seperti ringkasan, penyusunan draft, penataan poin, dan penjelasan awal terasa sangat mudah. Agentic AI membuat workflow multi-langkah tampak dapat “diurus” oleh sistem. Dari sini, organisasi bisa tergoda memperluas peran AI terlalu cepat, misalnya:

- menjadikan draft AI hampir selalu sebagai naskah final;
- membiarkan AI menentukan prioritas tanpa pemeriksaan yang cukup;
- mengandalkan ringkasan AI sebagai satu-satunya dasar orientasi;
- atau membiarkan agent menjalankan urutan kerja tanpa jeda untuk evaluasi manusia.

Over-automation berbahaya karena dapat:

- menurunkan ketelitian manusia;
- mengaburkan tanggung jawab;
- membuat kesalahan AI lebih mudah menyebar ke langkah berikutnya;
- dan melemahkan kemampuan organisasi untuk mengenali kapan sistem sebenarnya sedang keliru.

Dalam sektor publik, over-automation juga berbahaya karena proses kerja seringkali berkaitan dengan:

- kewenangan formal;
- pertimbangan substantif;
- interpretasi kebijakan;
- atau pelayanan kepada masyarakat.

Semua ini tidak boleh diserahkan terlalu jauh kepada AI tanpa pembatasan.

Penting untuk dipahami bahwa over-automation tidak selalu muncul dalam bentuk yang ekstrim. Kadang ia hadir pelan-pelan, misalnya ketika:

- pengguna makin jarang membaca sumber asli karena merasa ringkasan AI sudah cukup;
- petugas makin jarang mengoreksi usulan kategori;
- atau pimpinan mulai menganggap bahan AI sebagai representasi penuh dari situasi yang sebenarnya lebih kompleks.

Karena itu, organisasi perlu menjaga agar AI tetap diposisikan sebagai **bantuan**, bukan sebagai pengganti tanggung jawab. Salah satu cara utama menghindari over-automation adalah dengan merancang *human-in-the-loop* yang nyata, bukan hanya asumsi bahwa manusia “masih ada di sana”.

5.6.5 Tindakan agent yang tidak terkendali

Risiko kelima adalah **tindakan agent yang tidak terkendali**. Ini adalah risiko yang sangat khas pada agentic AI, karena agent tidak hanya menghasilkan keluaran, tetapi dapat menjalankan beberapa langkah atau memicu tindakan tertentu dalam workflow.

Tindakan yang tidak terkendali bisa muncul bila:

- tujuan yang diberikan terlalu luas;
- batas kewenangan agent tidak jelas;
- agent memiliki akses ke terlalu banyak tool atau sumber;
- tidak ada titik persetujuan manusia;
- atau sistem tidak memiliki guardrail yang cukup untuk menentukan kapan agent harus berhenti.

Sebagai contoh, agent dapat:

- mengambil sumber yang tidak tepat lalu tetap menyusun jawaban seolah-olah sah;
- memilih langkah berikutnya yang sebenarnya tidak diinginkan organisasi;
- menggabungkan beberapa fungsi secara otomatis tanpa konteks yang cukup;
- atau mendorong workflow ke arah yang terlalu jauh sebelum manusia sempat memeriksa hasil antara.

Bahaya utamanya adalah bahwa tindakan agent yang tidak terkendali bisa menciptakan **kesalahan berantai**. Bila langkah awal salah, maka langkah-langkah berikutnya juga berangkat dari dasar yang salah. Karena agentic AI tampak “efisien”, organisasi kadang baru menyadari masalah ketika kesalahan itu sudah sampai ke keluaran akhir atau bahkan mempengaruhi proses kerja.

Dalam konteks pemerintahan, risiko ini sangat serius karena:

- alur kerja organisasi mengandung kewenangan dan tanggung jawab;
- beberapa langkah tidak boleh dilakukan tanpa persetujuan;
- dan tindakan yang keliru dapat berdampak pada layanan, dokumen, atau pengetahuan organisasi.

Karena itu, agentic AI perlu diintegrasikan ke workflow secara jauh lebih hati-hati dibanding generative AI biasa. Organisasi perlu menetapkan:

- tindakan apa yang boleh dilakukan agent;
- sumber apa yang boleh diakses;
- kapan agent harus berhenti dan meminta persetujuan;
- log apa yang harus dicatat;
- dan bagaimana fallback dilakukan bila agent menemui kasus yang tidak jelas.

Dengan kata lain, agentic AI tidak boleh diberi kebebasan hanya karena secara teknis mampu menjalankan beberapa langkah.

Hal yang perlu dipahami adalah bahwa kelima risiko di atas tidak berdiri sendiri. Mereka dapat saling memperkuat. Misalnya:

- prompt injection dapat menyebabkan data leakage;
- data leakage dapat muncul melalui output yang confabulatory;
- confabulation yang tidak diperiksa dapat memperkuat over-automation;
- over-automation dapat membuat tindakan agent yang tidak terkendali lebih sulit dihentikan;
- dan tindakan agent yang tidak terkendali dapat memperluas paparan data sensitif atau memperparah kesalahan informasi.

Karena itu, organisasi tidak cukup mengelola setiap risiko secara terpisah. Yang dibutuhkan adalah desain sistem yang menyatukan:

- pembatasan data,
- pembatasan prompt,
- kontrol tindakan agent,
- logging,
- pemeriksaan manusia,
- dan evaluasi penggunaan secara berkelanjutan.

Bagi organisasi, kelima risiko ini dapat diterjemahkan ke dalam beberapa pertanyaan praktis:

- Apakah sistem ini bisa menghasilkan jawaban atau ringkasan yang tampak meyakinkan tetapi tidak didukung sumber?
- Apakah input pengguna atau dokumen yang dibaca sistem bisa memanipulasi perilaku model?
- Apakah ada kemungkinan data sensitif muncul kembali melalui output?
- Apakah pengguna mulai terlalu percaya pada keluaran AI?
- Apakah agent diberi ruang gerak yang lebih luas daripada yang dibutuhkan?
- Apakah ada titik persetujuan manusia yang cukup jelas?
- Apakah log dan jejak langkah agent cukup tersedia untuk evaluasi?

Pertanyaan-pertanyaan seperti ini membantu organisasi menilai apakah workflow berbasis generative AI atau agentic AI masih berada dalam batas yang sehat.

Semua organisasi perlu memperhatikan risiko generative AI dan agentic AI, tetapi sektor publik punya alasan tambahan untuk lebih berhati-hati. Ini karena banyak proses pemerintahan menyangkut:

- layanan kepada warga;
- dokumen resmi;
- data sensitif;
- interpretasi aturan;
- atau pengambilan keputusan yang berdampak lebih luas daripada sekadar efisiensi internal.

Dalam konteks seperti ini, kesalahan AI bukan hanya masalah teknis. Ia bisa menjadi masalah tata kelola, keadilan, privasi, kepercayaan publik, dan legitimasi organisasi. Karena itu, sektor publik tidak cukup hanya “mencoba teknologi baru”, tetapi harus menempatkannya dalam kerangka pengawasan, etika, dan tanggung jawab yang lebih kuat.

Generative AI dan agentic AI membawa risiko-risiko yang lebih khas dibanding banyak bentuk AI lainnya. Confabulation membuat sistem tampak meyakinkan padahal keliru. Prompt injection membuka ruang manipulasi terhadap perilaku sistem. Data leakage membuat informasi sensitif dapat keluar melalui input, output, atau integrasi yang tidak terkendali. Over-automation mendorong organisasi terlalu

cepat menyerahkan proses ke AI. Tindakan agent yang tidak terkendali dapat memperluas kesalahan dan mengaburkan akuntabilitas.

5.7 Prinsip etika AI dan nilai publik

Setelah membahas keamanan AI, risiko pada data, model, output, prompt, dan vendor, serta risiko khusus pada generative AI dan agentic AI, langkah berikutnya adalah menempatkan seluruh pembahasan tersebut di dalam kerangka yang lebih mendasar, yaitu **prinsip etika AI dan nilai publik**. Bagian ini penting karena keamanan dan pengendalian risiko, walaupun sangat penting, belum cukup dengan sendirinya. Sebuah sistem AI bisa saja aman secara teknis, tetapi tetap bermasalah bila tidak adil, tidak menghormati privasi, tidak dapat dijelaskan secara memadai, atau dipakai dengan cara yang tidak selaras dengan kepentingan publik. Karena itu, organisasi publik perlu mempunyai landasan etika yang cukup jelas agar pemanfaatan AI tidak hanya efektif, tetapi juga patut dan dapat dipertanggungjawabkan.

Peta Jalan Kecerdasan Artifisial Nasional Tahun 2026–2029 secara tegas menetapkan bahwa pelaksanaan AI perlu berpedoman pada nilai-nilai Pancasila dan UUD 1945, serta prinsip kebermanfaatan, bermartabat, keadilan, akuntabilitas, perlindungan data pribadi, transparansi, keamanan, keberlanjutan, integritas, inklusivitas, serta keterlibatan dan pengawasan manusia. Dokumen yang sama juga menempatkan mitigasi risiko sebagai salah satu arah kebijakan. Ini menunjukkan bahwa dalam konteks Indonesia, etika AI bukan pelengkap setelah teknologi jadi, melainkan bagian dari arah kebijakan nasional.

Hal yang sejalan juga terlihat dalam **Konsep Pedoman Etika Kecerdasan Artifisial**. Dokumen tersebut menegaskan bahwa etika diperlukan untuk memastikan pengembangan, penerapan, dan pemanfaatan AI dilakukan secara bertanggung jawab, adil, transparan, menghormati privasi, menjamin keamanan, dan memungkinkan pengawasan manusia. Dokumen itu juga menempatkan langkah-langkah keselamatan, keamanan, identifikasi dan mitigasi risiko, peran dan tanggung jawab para pihak, serta pemantauan dan evaluasi sebagai bagian dari kerangka etika AI.

Dari sini dapat dipahami bahwa prinsip etika AI dalam sektor publik tidak berdiri di luar penggunaan AI, melainkan harus menjiwai seluruh siklus hidupnya: mulai dari memilih use case, merancang solusi, menentukan data, memilih model, mengintegrasikan ke workflow, sampai memantau hasil dan memperbaikinya. Prinsip etika membantu organisasi menjawab pertanyaan yang lebih dalam daripada sekadar “bisakah AI dipakai di sini?” Pertanyaan yang perlu dijawab adalah:

- apakah penggunaan AI ini memberi manfaat yang sah dan layak?
- apakah ia memperlakukan pengguna dan pihak terdampak secara adil?
- apakah ia menjaga martabat manusia?
- apakah risikonya dipahami dan dikendalikan?
- apakah hasilnya dapat dijelaskan dan dipertanggungjawabkan?
- dan apakah penggunaan AI ini benar-benar mendukung nilai publik?

5.7.1 Etika AI

Seringkali etika dipahami secara sempit sebagai kepatuhan terhadap aturan atau sebagai daftar larangan. Dalam konteks AI, cara pandang seperti itu belum cukup. Etika AI lebih tepat dipahami sebagai **kerangka penilaian dan pengambilan keputusan** yang membantu organisasi menentukan:

- apa yang patut dilakukan;
- apa yang harus dibatasi;
- bagaimana menyeimbangkan manfaat dan risiko;
- dan bagaimana memastikan bahwa inovasi tidak berjalan dengan mengorbankan hak, keadilan, dan kepercayaan.

Dengan kata lain, etika AI bukan hanya tentang menghindari kesalahan, tetapi juga tentang memastikan bahwa organisasi menggunakan AI dengan arah yang benar. Dalam sektor publik, hal ini penting karena organisasi tidak bekerja hanya untuk dirinya sendiri, melainkan untuk masyarakat, tata kelola yang sehat, dan keberlanjutan institusi.

1. Prinsip kebermanfaatan

Prinsip pertama yang sangat mendasar adalah **kebermanfaatan**. AI seharusnya digunakan untuk memberi manfaat yang nyata dan sah, bukan semata-mata karena teknologinya tersedia atau sedang populer. Prinsip ini menuntut organisasi untuk selalu kembali pada pertanyaan:

- masalah apa yang sedang dibantu;
- siapa yang menerima manfaatnya;
- dan apakah manfaat tersebut cukup jelas untuk membenarkan penggunaan AI.

Dalam sektor publik, kebermanfaatan tidak hanya berarti efisiensi internal. Kebermanfaatan juga bisa berarti:

- akses informasi yang lebih baik;
- layanan yang lebih responsif;
- orientasi kebijakan yang lebih cepat;
- penguatan pengawasan;
- atau pengurangan beban kerja repetitif agar pegawai dapat lebih fokus pada tugas yang memerlukan penilaian manusia.

Namun, kebermanfaatan perlu dibaca secara proporsional. Tidak semua use case yang bisa dibuat benar-benar bermanfaat dalam arti publik. Karena itu, prinsip ini mendorong organisasi memilih penggunaan AI yang benar-benar punya alasan yang kuat.

2. Prinsip keadilan dan fairness

Prinsip kedua adalah **keadilan**. Dalam AI, prinsip ini berkaitan dengan kewaspadaan terhadap bias, ketimpangan, dan hasil yang dapat merugikan kelompok, konteks, atau jenis kasus tertentu secara tidak proporsional. Dalam sektor publik, prinsip keadilan sangat penting karena penggunaan AI dapat mempengaruhi bagaimana organisasi:

- mengelompokkan informasi,
- memberi prioritas,
- menyediakan jawaban,
- atau membentuk orientasi awal terhadap masalah.

Keadilan tidak selalu berarti semua orang diperlakukan identik tanpa konteks. Yang lebih penting adalah bahwa AI tidak menciptakan atau memperkuat perlakuan yang tidak semestinya. Organisasi perlu peka terhadap pertanyaan seperti:

- apakah sistem lebih sering salah pada jenis input tertentu;
- apakah ada kelompok yang lebih sering dirugikan oleh pola hasil AI;
- apakah hasil AI cenderung memusatkan perhatian hanya pada kasus yang dominan dan mengabaikan yang minor tetapi penting.

Prinsip keadilan membantu organisasi menjaga agar AI tidak menjadi alat yang mempercepat ketidakadilan yang sudah ada.

1. Prinsip penghormatan terhadap martabat manusia

Dokumen etika AI nasional menegaskan bahwa pedoman etika AI harus selaras dengan nilai Pancasila, hak asasi manusia, dan penghormatan terhadap martabat manusia. Prinsip ini penting karena AI tidak boleh membuat manusia hanya diperlakukan sebagai objek data atau sekadar titik dalam sistem.

Dalam praktik organisasi, penghormatan terhadap martabat manusia berarti antara lain:

- tidak menggunakan AI dengan cara yang merendahkan pengguna atau pihak terdampak;
- tidak menghapus ruang bagi pertimbangan manusia pada konteks yang sensitif;
- tidak menjadikan individu sekadar bahan otomatisasi tanpa hak untuk diperlakukan secara layak;
- dan tidak membiarkan AI mendorong perlakuan yang kaku pada situasi yang sebenarnya memerlukan konteks dan empati.

Prinsip ini sangat penting terutama pada layanan publik, pengaduan, komunikasi kepada warga, dan proses yang melibatkan data pribadi atau kondisi sensitif.

3. Prinsip transparansi

Prinsip berikutnya adalah **transparansi**. Dalam konteks AI, transparansi bukan berarti semua detail teknis harus dibuka ke semua orang, tetapi bahwa organisasi perlu cukup jelas mengenai:

- kapan AI digunakan;
- untuk tujuan apa AI digunakan;
- data atau sumber apa yang menjadi dasar;
- apa batas penggunaan sistem;
- dan bagaimana hasil AI seharusnya dibaca.

Konsep Pedoman Etika AI juga menekankan bahwa transparansi membantu membangun kepercayaan. Bahkan pada lampiran dokumen tersebut muncul gagasan “transparency builds trustworthiness”. Transparansi penting karena AI yang tidak jelas cara penggunaannya mudah menimbulkan rasa tidak aman, salah paham, atau kepercayaan yang berlebihan.

Dalam organisasi publik, transparansi juga penting untuk audit, evaluasi, dan komunikasi internal. Pengguna perlu tahu apakah mereka sedang membaca hasil AI, apakah itu hanya bantuan awal, dan dari mana rujukannya.

4. Prinsip akuntabilitas

Prinsip selanjutnya adalah **akuntabilitas**. AI tidak boleh membuat tanggung jawab menjadi kabur. Ketika sistem AI digunakan dalam proses organisasi, harus tetap jelas:

- siapa yang merancang penggunaannya;
- siapa yang memeriksa hasilnya;
- siapa yang menyetujui tindak lanjut;
- dan siapa yang bertanggung jawab bila terjadi kesalahan.

Peta Jalan AI Nasional secara eksplisit memasukkan akuntabilitas sebagai prinsip pelaksanaan AI. Ini penting karena dalam banyak kasus, AI justru berisiko membuat orang berlindung di balik sistem. Misalnya, keputusan awal dianggap “sudah dari AI”, sehingga tanggung jawab manusia terasa berkurang. Cara pandang seperti ini harus dicegah.

Akuntabilitas berarti AI tetap diposisikan sebagai alat bantu, sedangkan tanggung jawab kelembagaan tetap berada pada manusia dan organisasi.

5. Prinsip perlindungan data pribadi dan privasi

Prinsip berikutnya adalah **perlindungan data pribadi dan privasi**. Prinsip ini sudah dibahas lebih khusus pada subbagian sebelumnya, tetapi dari sudut etika, ia perlu dipahami sebagai penghormatan terhadap batas-batas informasi pribadi dan informasi sensitif.

Dalam praktik, prinsip ini menuntut organisasi untuk:

- tidak memproses data secara berlebihan;
- membatasi akses;
- hanya memakai data yang benar-benar diperlukan;
- menjelaskan tujuan penggunaan data;
- dan menjaga agar AI tidak menjadi saluran kebocoran atau pengungkapan yang tidak semestinya.

Privasi bukan hanya isu teknis atau kepatuhan, tetapi bagian dari penghormatan terhadap hak dan kepercayaan.

6. Prinsip keamanan dan keselamatan

Konsep Pedoman Etika AI membedakan **keselamatan** (*safety*) dan **keamanan** (*security*). Dokumen tersebut menjelaskan bahwa keselamatan berkaitan dengan kemampuan sistem AI melindungi pengguna, memastikan pengembangan yang bertanggung jawab, handal, dan transparan dalam keputusan, sedangkan keamanan berkaitan dengan perlindungan dari serangan luar, pengelolaan akses, autentikasi, dan hal lain yang membangun sistem AI yang aman.

Dari sudut etika, keamanan dan keselamatan bukan hanya persoalan teknis. Keduanya adalah kewajiban moral dan kelembagaan agar AI tidak menimbulkan kerugian yang seharusnya bisa dicegah. Dalam organisasi publik, ini berarti:

- menjaga agar AI tidak mudah dimanipulasi;
- menjaga agar keluaran tidak membahayakan proses atau pengguna;
- dan memastikan bahwa sistem tidak dipakai melampaui tujuan yang dirancang.

7. Prinsip akuntabilitas

Peta Jalan AI Nasional juga menempatkan **integritas** sebagai salah satu prinsip. Dalam konteks AI, integritas dapat dibaca sebagai upaya menjaga agar:

- informasi yang dihasilkan tidak menyesatkan;
- sistem tidak dipakai untuk tujuan yang tidak semestinya;
- data dan sumber tetap dapat dipercaya;
- dan proses penggunaan AI tetap sesuai dengan nilai dan tujuan organisasi.

Prinsip ini penting terutama untuk melawan kecenderungan seperti confabulation, ringkasan yang merusak konteks, atau penggunaan AI yang terlalu bebas sehingga mutu pengetahuan organisasi menurun.

8. Prinsip inklusivitas

Prinsip berikutnya adalah **inklusivitas**. Ini berarti organisasi perlu peka terhadap siapa yang mungkin terdampak atau tertinggal dalam penggunaan AI. Dalam sektor publik, inklusivitas penting karena:

- tidak semua pengguna memiliki tingkat literasi digital yang sama;
- tidak semua jenis bahasa, gaya penulisan, atau bentuk dokumen akan diperlakukan setara oleh sistem;
- dan tidak semua unit atau kelompok memiliki kemampuan yang sama untuk beradaptasi terhadap workflow berbasis AI.

Inklusivitas menuntut organisasi untuk tidak menganggap AI sebagai solusi yang netral bagi semua orang. Organisasi perlu memperhatikan apakah ada pihak yang justru lebih dirugikan, lebih bingung, atau lebih sulit memperoleh manfaat dari sistem yang dibangun.

9. Prinsip keterlibatan dan pengawasan manusia

Prinsip **keterlibatan dan pengawasan manusia** juga sangat penting, dan secara eksplisit muncul dalam Peta Jalan AI Nasional maupun Konsep Pedoman Etika AI. Prinsip ini menegaskan bahwa AI tidak boleh bekerja tanpa kerangka pengawasan yang memadai, terutama pada konteks yang sensitif.

Dalam praktik, prinsip ini berarti:

- manusia menetapkan tujuan penggunaan AI;
- manusia memeriksa hasil pada titik-titik penting;
- manusia tetap memegang keputusan final bila diperlukan;
- dan organisasi memastikan bahwa AI tidak menghapus ruang untuk koreksi, pertimbangan, dan tanggung jawab manusia.

Prinsip ini sangat penting untuk menghindari over-automation dan tindakan agent yang tidak terkendali.

5.7.2 Nilai publik sebagai pusat etika AI sektor publik

Seluruh prinsip di atas pada akhirnya perlu diarahkan pada **nilai publik**. Dalam sektor publik, etika AI tidak cukup hanya dibaca sebagai kepatuhan pada norma abstrak. Etika perlu dihubungkan dengan pertanyaan apakah AI:

- memperkuat kualitas layanan;
- meningkatkan akses terhadap informasi yang benar;
- menjaga hak dan perlindungan pengguna;
- memperkuat tata kelola;
- dan mendukung kepercayaan masyarakat terhadap institusi.

Dengan kata lain, nilai publik adalah cara untuk menguji apakah prinsip etika tadi benar-benar hidup dalam praktik. Sebuah solusi AI mungkin efisien, tetapi bila merusak keadilan, privasi, atau akuntabilitas, maka nilainya bagi publik perlu dipertanyakan. Sebaliknya, solusi yang lebih sederhana tetapi lebih selaras dengan nilai publik justru bisa lebih layak diprioritaskan.

Satu hal penting adalah bahwa prinsip-prinsip etika ini tidak boleh berhenti sebagai daftar nilai yang indah. Organisasi perlu menerjemahkannya ke dalam praktik, misalnya:

- memilih use case yang benar-benar bermanfaat dan layak;
- membatasi data yang dipakai;
- menyiapkan dokumentasi dan sumber rujukan;
- menguji sistem pada variasi kasus yang relevan;
- menetapkan titik pemeriksaan manusia;
- membangun kebijakan penggunaan internal;
- dan meninjau hasil AI secara berkala.

Konsep Pedoman Etika AI sendiri menekankan bahwa etika AI tidak hanya berbicara tentang nilai, tetapi juga tentang manfaat penerapan nilai etika, identifikasi dan mitigasi risiko, langkah-langkah perlindungan, peran dan tanggung jawab, serta pemantauan dan evaluasi. Ini menunjukkan bahwa etika perlu hadir dalam bentuk kerja yang nyata.

Pada akhirnya, prinsip etika AI membantu organisasi menjawab pertanyaan-pertanyaan yang tidak selalu dapat dijawab hanya dengan logika teknis, misalnya:

- meskipun sesuatu bisa dibuat, apakah patut dipakai?
- meskipun sesuatu efisien, apakah cukup adil?
- meskipun modelnya kuat, apakah cukup dapat diawasi?
- meskipun hasilnya cepat, apakah cukup menjaga integritas informasi?
- dan meskipun vendor menawarkan kemudahan, apakah organisasi tetap bisa mempertahankan tanggung jawab dan kendalinya?

Pertanyaan-pertanyaan seperti ini penting karena penggunaan AI di sektor publik selalu mengandung dimensi normatif. Itulah sebabnya prinsip etika perlu menjadi dasar pengambilan keputusan, bukan sekadar pelengkap setelah keputusan diambil.

Prinsip etika AI dan nilai publik adalah landasan yang membantu organisasi publik memastikan bahwa AI digunakan secara patut, bertanggung jawab, dan selaras dengan tujuan kelembagaan. Prinsip-prinsip seperti kebermanfaatan, keadilan, penghormatan terhadap martabat manusia, transparansi, akuntabilitas, perlindungan data pribadi, keamanan, keselamatan, integritas, inklusivitas, serta keterlibatan dan pengawasan manusia bukan hanya nilai umum, tetapi arah praktis bagi desain, penggunaan, dan evaluasi AI.

5.8 Human oversight dan safeguard

Pada subbagian sebelumnya telah dibahas bahwa penggunaan AI di sektor publik perlu berlandaskan prinsip etika dan nilai publik, seperti kebermanfaatan, keadilan, transparansi, akuntabilitas, perlindungan data pribadi, integritas, inklusivitas, dan keterlibatan manusia. Prinsip-prinsip tersebut memberi arah normatif tentang bagaimana AI seharusnya diposisikan. Namun, agar prinsip itu tidak berhenti sebagai pernyataan yang baik di atas kertas, organisasi perlu menerjemahkannya ke dalam bentuk yang lebih operasional. Di sinilah pembahasan tentang **human oversight** dan **safeguard** menjadi sangat penting.

Secara sederhana, **human oversight** berarti adanya keterlibatan dan pengawasan manusia yang nyata terhadap penggunaan AI. Sementara itu, **safeguard** berarti berbagai batas, kontrol, pelindung, dan mekanisme pencegahan yang dirancang agar AI tidak melampaui ruang penggunaan yang semestinya, tidak menimbulkan kerugian yang tidak perlu, dan tidak mengaburkan tanggung jawab organisasi. Kedua hal ini saling terkait. Pengawasan manusia tanpa safeguard yang memadai akan terlalu bergantung pada kewaspadaan individu. Sebaliknya, safeguard teknis tanpa pengawasan manusia dapat membuat organisasi terlalu percaya bahwa sistem akan selalu berada dalam jalur yang benar.

Dalam **Peta Jalan Kecerdasan Artifisial Nasional Tahun 2026–2029**, keterlibatan dan pengawasan manusia ditempatkan sebagai salah satu prinsip pelaksanaan AI. Dokumen tersebut juga menempatkan mitigasi risiko, akuntabilitas, perlindungan data pribadi, keamanan, integritas, dan keadilan sebagai prinsip yang berjalan bersama. Ini menunjukkan bahwa pengawasan manusia tidak dipahami sebagai pelengkap setelah sistem jadi, tetapi sebagai bagian dari cara AI seharusnya digunakan. Hal yang sejalan juga tampak dalam **Konsep Pedoman Etika Kecerdasan Artifisial**, yang menegaskan pentingnya transparansi, keamanan, pengawasan manusia, identifikasi dan mitigasi risiko, serta kejelasan peran dan tanggung jawab para pihak dalam seluruh siklus hidup AI. Dengan kata lain, human oversight dan safeguard merupakan jembatan antara prinsip etika dan praktik tata kelola sehari-hari.

Ada kecenderungan untuk berpikir bahwa semakin canggih AI, semakin kecil kebutuhan akan campur tangan manusia. Dalam sektor publik, cara pandang seperti ini justru berbahaya. AI dapat membantu manusia bekerja lebih cepat, lebih terarah, dan lebih ringan, tetapi AI tidak menghapus kebutuhan akan pertimbangan, akuntabilitas, dan tanggung jawab manusia.

Human oversight diperlukan karena beberapa alasan utama.

- Pertama, AI tetap dapat menghasilkan keluaran yang salah, tidak lengkap, tidak adil, atau tidak sesuai konteks. Ini bisa terjadi pada ringkasan, klasifikasi, retrieval, draft awal, maupun workflow agentic. Tanpa pengawasan manusia, keluaran tersebut dapat langsung masuk ke proses kerja seolah-olah sah.
- Kedua, AI tidak memikul tanggung jawab kelembagaan. Dalam organisasi pemerintah, tanggung jawab tetap berada pada manusia dan institusi, bukan pada model, aplikasi, atau vendor. Karena itu, perlu selalu jelas siapa yang memeriksa, siapa yang menyetujui, dan siapa yang bertanggung jawab atas penggunaan hasil AI.
- Ketiga, AI tidak memahami konteks sosial, administratif, hukum, dan moral sebagaimana manusia. AI dapat membantu memproses informasi, tetapi ia tidak memiliki legitimasi untuk menafsirkan konsekuensi kelembagaan secara mandiri.
- Keempat, pengawasan manusia membantu organisasi belajar. Koreksi, keberatan, dan penilaian pengguna terhadap hasil AI memberi umpan balik yang penting untuk memperbaiki sistem, mengubah batas penggunaan, atau bahkan menghentikan penggunaan pada konteks tertentu.

Karena itu, human oversight tidak boleh dipahami sebagai tanda kelemahan AI. Justru, dalam sektor publik, human oversight adalah syarat agar manfaat AI dapat diperoleh tanpa kehilangan kendali organisasi.

5.8.1 Bentuk-bentuk human oversight

Dalam praktik organisasi, human oversight dapat muncul dalam beberapa bentuk.

Bentuk pertama adalah **oversight pada tahap tujuan dan desain**. Di sini manusia menentukan:

- masalah apa yang layak dibantu AI;
- tujuan penggunaannya;
- sumber data yang dipakai;
- batas ruang lingkup;
- dan tingkat risiko yang dapat diterima.

Bentuk kedua adalah **oversight pada tahap penggunaan operasional**. Ini adalah bentuk yang paling sering terlihat, misalnya:

- memeriksa hasil ringkasan;
- memvalidasi usulan kategori;
- menilai relevansi jawaban berbasis dokumen;
- menyunting draft awal;
- atau menilai apakah sinyal anomali memang perlu ditindaklanjuti.

Bentuk ketiga adalah **oversight pada tahap eskalasi dan persetujuan**. Dalam bentuk ini, AI dapat membantu beberapa langkah, tetapi manusia tetap harus menyetujui sebelum keluaran dipakai untuk langkah berikutnya atau dipakai sebagai dasar tindakan yang lebih formal.

Bentuk keempat adalah **oversight pada tahap evaluasi dan perbaikan**. Organisasi meninjau log, koreksi pengguna, hasil pengujian, insiden, atau pola kegagalan untuk melihat apakah sistem masih layak digunakan, perlu diperbaiki, atau perlu dibatasi lebih lanjut.

Dengan demikian, human oversight sesungguhnya hadir di seluruh siklus hidup AI, bukan hanya saat pengguna melihat output akhir.

Namun, tidak semua use case memerlukan tingkat pengawasan manusia yang sama. Salah satu prinsip penting adalah bahwa **tingkat oversight harus sebanding dengan tingkat risiko**.

- Untuk use case berisiko rendah, misalnya ringkasan bahan internal untuk orientasi awal, pengawasan manusia bisa relatif ringan. Pengguna cukup memeriksa apakah isi ringkasan masih sesuai dengan sumber dan cukup berguna untuk pekerjaannya.
- Untuk use case berisiko menengah, seperti klasifikasi awal aduan atau drafting awal tanggapan internal, pengawasan perlu lebih jelas dan lebih konsisten. Pengguna harus paham bahwa hasil AI adalah usulan atau bahan awal, bukan hasil final.
- Untuk use case yang lebih sensitif, seperti yang melibatkan data pribadi, dokumen resmi, komunikasi ke publik, atau langkah yang dapat mempengaruhi hak dan kepentingan pihak lain, pengawasan manusia harus lebih kuat. Dalam konteks seperti ini, hasil AI sebaiknya tidak pernah dipakai tanpa verifikasi yang cukup.

Pendekatan berbasis risiko ini penting agar organisasi tidak menerapkan pola pengawasan yang terlalu longgar pada konteks sensitif, dan juga tidak membuat proses menjadi terlalu berat pada konteks yang sebenarnya sederhana.

5.8.2 Safeguard

Jika human oversight menekankan peran manusia, maka **safeguard** menekankan perangkat pelindung yang dibangun ke dalam sistem, proses, dan tata kelola. Safeguard dapat berupa:

- pembatasan akses,
- pembatasan sumber data,
- pembatasan tindakan agent,
- validasi input,
- guardrails pada prompt,
- pembatasan output,
- logging,
- notifikasi eskalasi,
- kontrol perubahan,
- sampai prosedur fallback ketika AI tidak boleh melanjutkan.

Safeguard penting karena organisasi tidak bisa hanya mengandalkan kewaspadaan manusia setiap saat. Sistem dan workflow perlu dirancang agar sejak awal tidak terlalu mudah tergelincir ke penggunaan yang salah. Dengan kata lain, safeguard adalah cara organisasi “membangun pagar” sebelum masalah terjadi.

1. Safeguard pada data dan akses

Salah satu bentuk safeguard yang paling mendasar adalah pembatasan pada **data dan akses**. Organisasi perlu memastikan bahwa AI:

- hanya mengakses data yang memang diperlukan;
- hanya membaca sumber yang relevan dengan use case;
- tidak diberi akses ke repositori yang terlalu luas tanpa alasan;
- dan tidak menampilkan informasi yang melebihi hak akses pengguna.

Ini sangat penting terutama pada knowledge assistant, retrieval system, generative AI berbasis dokumen, dan agentic AI yang terhubung ke beberapa tool atau sumber. Bila safeguard pada akses lemah, maka risiko kebocoran data, jawaban yang melampaui kewenangan, atau agent yang bergerak terlalu luas akan meningkat.

2. Safeguard pada prompt dan instruksi

Pada generative AI dan agentic AI, safeguard juga perlu diterapkan pada **prompt** dan instruksi. Organisasi perlu memikirkan:

- prompt seperti apa yang diperbolehkan untuk tugas tertentu;
- bagaimana system prompt atau guardrails dijaga;
- bagaimana input dari pengguna divalidasi atau dibatasi;
- bagaimana sistem membedakan antara konten dokumen dan instruksi kepada model;

- dan bagaimana agent harus bereaksi terhadap instruksi yang ambigu, manipulatif, atau bertentangan dengan tujuan sistem.

Safeguard seperti ini penting untuk menekan risiko prompt injection, keluaran yang terlalu bebas, serta penggunaan AI di luar batas tugasnya.

3. Safeguard pada output

Output AI juga memerlukan safeguard. Organisasi sebaiknya tidak membiarkan hasil AI tampil begitu saja tanpa pembimbingan yang memadai. Safeguard pada output bisa berupa:

- label yang jelas bahwa hasil adalah bantuan awal atau usulan;
- tampilan sumber rujukan;
- peringatan ketika keyakinan atau kelengkapan sumber rendah;
- pembatasan isi tertentu agar tidak keluar;
- serta mekanisme agar hasil tertentu harus ditinjau sebelum bisa dipakai lebih lanjut.

Ini sangat penting karena banyak risiko AI justru terjadi ketika output yang terlihat rapi dan meyakinkan diterima begitu saja tanpa penjelasan statusnya.

4. Safeguard pada workflow agentic

Pada agentic AI, safeguard harus lebih kuat lagi karena agent dapat membantu beberapa langkah sekaligus. Beberapa bentuk safeguard yang penting antara lain:

- membatasi tool atau fungsi apa yang boleh dipakai agent;
- membatasi sumber mana yang boleh diakses;
- menetapkan langkah mana yang wajib mendapat persetujuan manusia;
- menentukan kondisi kapan agent harus berhenti dan menyerahkan proses;
- mencatat seluruh langkah yang dijalankan;
- serta mencegah agent mengambil tindakan yang melebihi ruang tugasnya.

Tanpa safeguard seperti ini, agentic AI berisiko menjalankan workflow yang tampak efisien tetapi sulit diawasi dan mudah melampaui batas yang dirancang.

Human oversight dan safeguard tidak boleh dirancang terpisah. Keduanya harus saling mendukung. Misalnya:

- jika organisasi menetapkan bahwa hasil AI harus diperiksa manusia, maka aplikasi dan workflow harus menyediakan ruang koreksi yang mudah;
- jika organisasi menetapkan bahwa agent harus berhenti pada kondisi tertentu, maka sistem perlu memiliki trigger atau aturan untuk memaksa penghentian itu;
- jika hasil ringkasan harus ditelusuri ke sumber, maka antarmuka perlu menampilkan sumber tersebut secara cukup jelas.

Dengan demikian, pengawasan manusia tidak cukup hanya berupa kebijakan, dan safeguard teknis tidak cukup hanya berupa fitur. Keduanya harus bertemu di dalam desain workflow dan aplikasi.

5.8.3 Peran pelatihan dan kapasitas pengguna

Human oversight dan safeguard akan lemah bila pengguna tidak memahami perannya. Karena itu, organisasi perlu memberi perhatian pada:

- pelatihan penggunaan AI;
- pemahaman tentang batas sistem;
- cara membaca hasil AI;
- kapan harus memeriksa ulang;
- dan bagaimana melaporkan hasil yang bermasalah.

Ini penting karena dalam banyak kasus, AI tidak gagal semata-mata karena sistemnya buruk, tetapi karena pengguna:

- terlalu percaya pada hasil;
- tidak tahu bahwa output perlu diperiksa;
- atau tidak memahami bahwa AI hanya dirancang untuk konteks tertentu.

Pelatihan dan literasi penggunaan AI dengan demikian menjadi bagian dari safeguard kelembagaan.

5.8.4 Safeguard dan fallback

Satu unsur penting lain adalah **fallback**, yaitu apa yang dilakukan organisasi bila AI:

- gagal memproses input,
- memberi hasil yang tidak cukup layak,
- tidak menemukan sumber yang cukup jelas,
- atau tidak boleh melanjutkan karena konteksnya terlalu sensitif.

Fallback penting karena menunjukkan bahwa organisasi tidak memaksa AI bekerja pada semua situasi. Dalam banyak kasus, fallback yang sehat bisa berupa:

- pengalihan ke proses manual;
- pengalihan ke petugas;
- penundaan tindakan sampai ada verifikasi;
- atau pembatasan keluaran pada informasi minimal.

Pendekatan seperti ini membantu organisasi menjaga agar AI tetap menjadi alat bantu, bukan titik kegagalan tunggal. Human oversight dan safeguard juga perlu ditopang oleh:

- **logging** untuk melihat apa yang dilakukan sistem;
- **dokumentasi** untuk menjelaskan tujuan, batas, dan mekanisme kontrol;
- serta **review berkala** untuk melihat apakah kontrol yang dirancang masih efektif.

Tanpa logging, organisasi sulit menilai apakah safeguard benar-benar bekerja. Tanpa dokumentasi, kontrol yang dibuat mudah hilang atau berubah tanpa disadari. Tanpa review berkala, sistem yang awalnya cukup aman bisa perlahan menyimpang karena perubahan data, model, vendor, atau perilaku pengguna.

Pada akhirnya, desain human oversight dan safeguard perlu menjaga keseimbangan. Bila terlalu longgar, AI bisa melampaui batas dan memperbesar risiko. Bila terlalu kaku, manfaat AI bisa sulit dirasakan karena setiap langkah menjadi terlalu berat dan terlalu lambat.

Organisasi perlu mencari titik keseimbangan yang sesuai dengan:

- tingkat risiko use case;
- kemampuan pengguna;
- kebutuhan kecepatan kerja;
- sensitivitas data;
- dan kapasitas organisasi untuk memantau sistem.

Keseimbangan ini tidak selalu sama untuk semua use case. Karena itu, human oversight dan safeguard perlu dirancang secara kontekstual.

Human oversight dan safeguard adalah dua unsur penting yang menerjemahkan prinsip etika AI ke dalam bentuk pengendalian yang nyata. Human oversight memastikan bahwa manusia tetap menetapkan tujuan, memeriksa hasil, mengoreksi keluaran, menyetujui langkah penting, dan memegang tanggung jawab akhir. Safeguard memastikan bahwa sistem, data, prompt, output, dan workflow dibatasi serta dilindungi agar AI tidak bergerak melampaui ruang penggunaan yang sah.

5.9 Manajemen risiko AI

Setelah membahas keamanan AI, risiko pada data, model, output, prompt, dan vendor, privasi, bias, integritas informasi, serta pentingnya *human oversight* dan *safeguard*, langkah berikutnya adalah melihat bagaimana semua itu dikelola secara lebih sistematis melalui **manajemen risiko AI**. Bagian ini penting karena dalam praktik organisasi, risiko AI tidak akan cukup terkendali hanya dengan kewaspadaan umum atau dengan beberapa kontrol yang berdiri sendiri. Organisasi memerlukan cara berpikir yang lebih terstruktur: bagaimana risiko diidentifikasi, dinilai, diprioritaskan, dikendalikan, dipantau, dan ditinjau ulang sepanjang siklus hidup sistem.

Dalam konteks sektor publik, pendekatan seperti ini sangat penting. Pemerintah tidak dapat mengandalkan logika “coba dulu, nanti diperbaiki kalau ada masalah” untuk semua bentuk penggunaan AI. Pada beberapa konteks, kesalahan atau kegagalan AI dapat berdampak pada:

- mutu layanan;
- keamanan informasi;
- akuntabilitas proses;
- kepercayaan publik;
- atau kualitas pengambilan keputusan awal.

Karena itu, AI perlu diperlakukan sebagai sesuatu yang risikonya dikelola sejak awal, bukan hanya diperiksa setelah terjadi insiden.

ISO/IEC 22989 menegaskan bahwa sistem AI memerlukan pendekatan **risk management** yang berkelanjutan dan menyeluruh pada setiap fase siklus hidupnya, termasuk terkait privasi, keselamatan, keamanan digital, dan bias. Standar ini juga menekankan bahwa AI system life cycle bersifat lebih iteratif dan kurang dapat diprediksi dibanding sistem non-AI biasa, sehingga memerlukan *continuous validation*, *re-evaluation*, dan pemantauan risiko sepanjang operasi.

Hal yang serupa ditegaskan dalam **Konsep Pedoman Etika Kecerdasan Artifisial**, yang menyebut pentingnya identifikasi dan mitigasi risiko, langkah-langkah perlindungan, peran dan tanggung jawab para pihak, serta pemantauan dan evaluasi risiko AI. Dokumen tersebut juga menekankan bahwa pemangku kebijakan dan pelaku sektor perlu menjalankan manajemen risiko di setiap tahap siklus hidup AI dan membangun mekanisme penilaian mandiri, pelaporan insiden, serta pemantauan terpadu.

Sementara itu, **Peta Jalan Kecerdasan Artifisial Nasional Tahun 2026–2029** juga menempatkan mitigasi risiko sebagai salah satu prinsip dan arah pelaksanaan AI nasional, bersama dengan akuntabilitas, keadilan, perlindungan data pribadi, integritas, keamanan, transparansi, dan keterlibatan manusia.

Dari dasar-dasar tersebut, manajemen risiko AI dapat dipahami sebagai **proses organisasi yang berkelanjutan** untuk menjaga agar penggunaan AI tetap berada dalam batas manfaat yang dapat diterima dan risiko yang dapat dikendalikan. Bukan untuk menghapus seluruh risiko—karena itu tidak realistis—melainkan untuk:

- mengenali risiko sejak awal;
- memahami konteks dan dampaknya;
- memilih kontrol yang sesuai;
- dan memantau apakah risiko itu tetap terkendali seiring sistem dipakai dan berubah.

Organisasi publik umumnya sudah memiliki kerangka umum untuk manajemen risiko, baik pada tingkat proyek, sistem informasi, keamanan siber, maupun operasional. Namun, AI membawa beberapa karakter yang membuat manajemen risikonya memerlukan perhatian tambahan.

1. **Pertama**, risiko AI tidak hanya terletak pada ketersediaan sistem atau kerahasiaan data, tetapi juga pada:
 - kualitas dan perilaku model,
 - kualitas output,
 - fairness,
 - perubahan perilaku karena prompt atau data baru,
 - serta cara manusia menafsirkan keluaran AI.
2. **Kedua**, AI sering menghasilkan keluaran yang tampak meyakinkan walaupun salah. Ini membuat risikonya tidak selalu mudah dideteksi seperti bug teknis biasa.
3. **Ketiga**, sistem AI bisa berubah perilakunya seiring perubahan data, model, vendor, prompt, atau workflow. Karena itu, risikonya tidak statis.
4. **Keempat**, risiko AI sering berada di perbatasan antara teknis, operasional, etika, dan tata kelola. Sebagai contoh:
 - kebocoran data adalah risiko keamanan dan privasi;
 - bias adalah risiko etika dan kualitas;
 - over-automation adalah risiko operasional dan akuntabilitas;
 - prompt injection adalah risiko teknis yang berdampak pada integritas proses;
 - dan perubahan model vendor adalah risiko pihak ketiga yang berdampak pada seluruh sistem.

Dengan kata lain, manajemen risiko AI perlu lebih lintas fungsi dibanding manajemen risiko TI biasa.

5.9.1 Risk register AI

Salah satu alat praktis yang sangat berguna adalah **risk register AI**, yaitu daftar risiko yang terdokumentasi untuk tiap use case atau sistem AI. Risk register dapat memuat hal-hal seperti:

- deskripsi risiko;
- kategori risiko;
- sumber risiko;
- dampak yang mungkin timbul;
- tingkat kemungkinan dan dampak;
- kontrol yang sudah ada;
- kontrol tambahan yang diperlukan;
- pihak yang bertanggung jawab;
- status penanganan;
- dan waktu peninjauan ulang.

Risk register membantu organisasi menjaga agar diskusi risiko tidak menghilang setelah fase awal proyek. Ia juga membantu komunikasi lintas fungsi, karena tim teknis, pemilik proses, pengelola data, dan pihak tata kelola dapat memakai rujukan yang sama.

Dalam manajemen risiko AI, penting juga menetapkan **pemilik risiko**. Tidak semua risiko harus dikelola oleh tim teknis. Beberapa risiko lebih tepat dimiliki oleh:

- pemilik proses bisnis;
- pengelola data;
- tim keamanan informasi;
- pengelola vendor;
- atau pimpinan unit yang memakai sistem.

Sebagai contoh:

- risiko bias pada klasifikasi aduan mungkin perlu dimiliki bersama oleh pemilik proses layanan dan tim pengelola data;
- risiko data leakage dari vendor perlu melibatkan pengelola vendor dan keamanan informasi;
- risiko over-automation pada workflow drafting perlu melibatkan pemilik proses yang memutuskan kapan AI boleh dipakai;
- risiko integritas ringkasan bahan kebijakan perlu melibatkan analis atau pemilik substansi.

Penetapan pemilik risiko penting agar tidak semua persoalan dilempar ke tim teknis, padahal banyak risiko AI sesungguhnya bersifat lintas fungsi.

Walaupun kontrol sudah diterapkan, biasanya tetap ada **risiko residual**, yaitu risiko yang masih tersisa. Ini penting dipahami agar organisasi tidak terjebak pada ilusi bahwa AI dapat dibuat sepenuhnya bebas risiko. Yang perlu diputuskan adalah:

- apakah risiko residual tersebut masih dapat diterima;
- apakah manfaat sistem cukup untuk membenarkannya;
- apakah perlu pembatasan tambahan;
- atau apakah use case perlu dipersempit atau dihentikan.

Di sinilah manajemen risiko AI menjadi keputusan organisasi, bukan hanya keputusan teknis. Pada beberapa konteks, organisasi mungkin memutuskan untuk tidak melanjutkan use case tertentu meskipun secara teknis bisa dibuat, karena risiko residualnya terlalu tinggi.

5.9.2 Manajemen risiko AI dan budaya organisasi

Pada akhirnya, manajemen risiko AI juga sangat dipengaruhi oleh **budaya organisasi**. Bila budaya organisasi terlalu berorientasi pada kecepatan dan demo, maka risiko cenderung diabaikan. Bila budaya organisasi terlalu takut mencoba hal baru, maka manfaat AI sulit muncul. Yang dibutuhkan adalah budaya yang:

- cukup berani untuk mencoba;
- cukup disiplin untuk menilai risiko;
- cukup jujur untuk mengakui keterbatasan sistem;
- dan cukup terbuka untuk memperbaiki desain penggunaan AI dari waktu ke waktu.

Dalam sektor publik, budaya seperti ini sangat penting agar AI tidak menjadi sekadar simbol inovasi, tetapi benar-benar menjadi alat yang dikelola secara matang.

Manajemen risiko AI adalah proses organisasi yang sistematis dan berkelanjutan untuk mengidentifikasi, menilai, mengendalikan, memantau, dan meninjau ulang risiko sepanjang siklus hidup penggunaan AI. Dalam sektor publik, pendekatan ini sangat penting karena risiko AI tidak hanya bersifat teknis, tetapi juga menyangkut layanan, privasi, fairness, integritas informasi, akuntabilitas, dan kepercayaan publik.

5.10 Tata kelola AI dalam organisasi pemerintah

Setelah membahas keamanan AI, risiko pada data, model, output, prompt, dan vendor, privasi, bias, integritas informasi, *human oversight*, *safeguard*, serta manajemen risiko AI, langkah berikutnya adalah melihat semua itu dalam kerangka yang lebih luas, yaitu **tata kelola AI dalam organisasi pemerintah**. Bagian ini sangat penting karena pada akhirnya penggunaan AI tidak hanya ditentukan oleh kualitas teknologi atau ketepatan use case, tetapi oleh kemampuan organisasi membangun aturan main, pembagian peran, mekanisme pengawasan, dan proses pengambilan keputusan yang jelas. Tanpa tata kelola yang memadai, AI mudah dipakai secara sporadis, tidak konsisten, sulit dievaluasi, dan rawan menimbulkan risiko yang sebenarnya bisa dicegah.

Dalam konteks organisasi pemerintah, tata kelola AI tidak boleh dipahami sebagai lapisan administratif tambahan yang datang setelah sistem selesai dibangun. Tata kelola justru perlu hadir sejak awal sebagai kerangka yang membantu organisasi menjawab pertanyaan-pertanyaan mendasar:

- siapa yang boleh mengusulkan penggunaan AI;
- siapa yang menilai kelayakannya;
- siapa yang menyetujui penggunaannya;
- siapa yang bertanggung jawab atas data, model, dan output;
- siapa yang memantau risiko dan insiden;
- dan bagaimana organisasi memutuskan apakah suatu sistem AI boleh diteruskan, dibatasi, diperbaiki, atau dihentikan.

Dengan kata lain, tata kelola AI adalah cara organisasi memastikan bahwa AI dipakai **secara sah, terarah, aman, etis, dan akuntabel**.

Setiap organisasi memerlukan tata kelola teknologi, tetapi organisasi pemerintah memiliki kebutuhan yang lebih kuat karena bekerja di bawah tuntutan:

- akuntabilitas publik;
- kepatuhan pada aturan;
- perlindungan hak warga;
- penggunaan sumber daya publik yang dapat dipertanggungjawabkan;
- serta kewajiban menjaga kepercayaan masyarakat.

AI dapat mempengaruhi:

- cara informasi diolah;
- cara kerja internal berubah;
- cara layanan dijalankan;
- dan cara keputusan awal dibentuk.

Karena itu, penggunaan AI tanpa tata kelola yang jelas berisiko menimbulkan beberapa persoalan sekaligus.

- Pertama, organisasi bisa memakai AI secara tidak konsisten. Satu unit mungkin memakai AI dengan pengawasan yang cukup, sementara unit lain memakainya tanpa kontrol yang memadai.
- Kedua, tanggung jawab bisa menjadi kabur. Ketika terjadi masalah, tidak jelas siapa yang harus menjawab: tim teknis, pemilik proses, pengguna, vendor, atau pimpinan.
- Ketiga, organisasi dapat kehilangan jejak keputusan. AI dipakai karena dianggap membantu, tetapi tidak ada catatan yang jelas tentang mengapa dipilih, apa batasnya, dan bagaimana risiko dinilai.
- Keempat, penggunaan AI bisa melebar ke area yang lebih sensitif tanpa pembahasan yang cukup, terutama bila tidak ada mekanisme persetujuan dan evaluasi yang jelas.

Salah satu hal yang perlu ditekankan sejak awal adalah bahwa tata kelola AI bukan hanya tanggung jawab tim teknis atau unit teknologi informasi. Memang, tim teknis memiliki peran penting dalam membangun, mengintegrasikan, menguji, dan memelihara sistem. Namun, banyak pertanyaan utama dalam AI justru bersifat lintas fungsi, misalnya:

- apakah use case ini layak dari sisi proses bisnis;
- apakah data yang dipakai patut dan relevan;
- apakah output AI akan dipakai untuk konteks yang sensitif;
- apakah pengguna memahami batas sistem;
- apakah vendor eksternal dapat diterima dari sisi risiko;
- dan apakah hasil sistem cukup selaras dengan nilai publik.

Karena itu, tata kelola AI perlu melibatkan beberapa pihak sekaligus, misalnya:

- pemilik proses bisnis;
- pengelola data;
- tim teknis;
- keamanan informasi;
- fungsi hukum atau kepatuhan bila relevan;
- pengawasan internal;
- dan pimpinan yang berwenang menetapkan arah organisasi.

Pendekatan lintas fungsi seperti ini penting agar tata kelola AI tidak terlalu sempit dan tidak terputus dari realitas kerja organisasi.

5.10.1 Tata kelola AI sebagai struktur keputusan

Secara praktis, tata kelola AI dapat dipahami sebagai **struktur keputusan**. Artinya, organisasi perlu cukup jelas tentang keputusan apa saja yang harus diambil, pada tahap apa, dan oleh siapa.

Beberapa jenis keputusan yang biasanya perlu diatur antara lain:

- keputusan untuk memulai atau menolak use case AI;
- keputusan tentang data dan sumber yang boleh dipakai;
- keputusan tentang strategi integrasi, termasuk build, buy, adapt, atau hybrid;
- keputusan tentang penggunaan vendor atau layanan pihak ketiga;
- keputusan tentang tingkat pengawasan manusia yang diperlukan;
- keputusan tentang apakah sistem cukup aman untuk dipakai terbatas atau lebih luas;
- keputusan tentang bagaimana insiden ditangani;
- serta keputusan tentang kapan sistem harus dievaluasi ulang, dibatasi, atau dihentikan.

Bila struktur keputusan ini tidak jelas, maka penggunaan AI mudah berjalan berdasarkan kebiasaan, antusiasme sesaat, atau keputusan informal yang tidak terdokumentasi dengan baik.

5.10.2 Unsur utama tata kelola AI dalam organisasi pemerintah

Agar lebih konkret, tata kelola AI dalam organisasi pemerintah dapat dilihat melalui beberapa unsur utama.

1. Kebijakan dan prinsip penggunaan

Organisasi perlu memiliki arah yang cukup jelas tentang bagaimana AI boleh digunakan. Ini tidak selalu berarti dokumen yang sangat panjang, tetapi setidaknya organisasi perlu memiliki prinsip dasar seperti:

- AI digunakan untuk tujuan yang sah dan relevan dengan mandat organisasi;
- AI tidak digunakan pada konteks tertentu tanpa pengawasan manusia;
- data yang dipakai harus sesuai dengan tujuan dan pembatasan yang jelas;
- hasil AI diposisikan sesuai tingkat risikonya;
- dan setiap penggunaan AI harus dapat dijelaskan serta dievaluasi.

Prinsip seperti ini menjadi dasar bagi keputusan-keputusan berikutnya.

2. Pembagian peran dan tanggung jawab

Tata kelola yang sehat memerlukan kejelasan tentang siapa melakukan apa. Misalnya:

- siapa pemilik use case;
- siapa pemilik data;
- siapa yang bertanggung jawab atas integrasi teknis;
- siapa yang memantau performa dan insiden;
- siapa yang menyetujui penggunaan di lapangan;
- dan siapa yang berhak menghentikan penggunaan bila risikonya tidak lagi dapat diterima.

Tanpa kejelasan ini, AI mudah menjadi sistem yang dipakai banyak orang tetapi tidak benar-benar dimiliki atau diawasi secara utuh.

3. Mekanisme penilaian dan persetujuan

Organisasi sebaiknya tidak membiarkan semua use case AI berjalan begitu saja tanpa penilaian. Perlu ada mekanisme untuk menilai:

- manfaat;
- risiko;
- sensitivitas data;
- kebutuhan pengawasan manusia;
- kesiapan proses;
- dan kelayakan vendor bila ada pihak ketiga yang terlibat.

Tidak semua use case perlu melewati tingkat penilaian yang sama. Use case berisiko rendah mungkin cukup melewati proses yang lebih sederhana. Namun, use case yang menyentuh data sensitif, layanan publik, atau workflow yang lebih penting perlu melalui penilaian yang lebih ketat.

4. Dokumentasi dan keterlacakan

Tata kelola AI memerlukan dokumentasi yang cukup tentang:

- tujuan use case;
- sumber data;
- model atau layanan yang digunakan;
- prompt dan guardrail yang relevan;
- hasil pengujian;
- batas penggunaan;
- serta perubahan penting sepanjang siklus hidup sistem.

Dokumentasi penting karena tata kelola tanpa jejak yang jelas akan sulit dijalankan secara konsisten dan sulit diaudit.

5. Monitoring, evaluasi, dan perbaikan

Tata kelola yang baik tidak berhenti pada persetujuan awal. Organisasi perlu memonitor:

- bagaimana sistem dipakai;
- pola kesalahan;
- koreksi pengguna;
- perubahan risiko;
- keluhan atau insiden;
- dan apakah manfaat yang diharapkan benar-benar muncul.

Dari sini, organisasi dapat memutuskan apakah sistem:

- tetap layak digunakan;
- perlu diperbaiki;
- perlu dibatasi ruang lingkupnya;
- atau sebaiknya dihentikan.

5.10.3 Hubungan antara tata kelola AI dan manajemen risiko

Tata kelola AI dan manajemen risiko saling berkaitan erat. Manajemen risiko membantu organisasi mengenali dan mengendalikan risiko secara lebih teknis dan operasional. Tata kelola memberikan struktur kelembagaan agar proses itu benar-benar dijalankan secara konsisten.

Dengan kata lain:

- manajemen risiko menjawab **risiko apa yang ada dan bagaimana mengendalikannya**;
- tata kelola menjawab **siapa yang bertanggung jawab memastikan semua itu dilakukan, dipantau, dan ditinjau ulang**.

Tanpa tata kelola, manajemen risiko mudah menjadi daftar yang tidak diikuti. Tanpa manajemen risiko, tata kelola mudah menjadi formalitas tanpa isi.

Tata kelola AI tidak selalu harus dimulai dalam bentuk yang sangat kompleks. Organisasi dapat membangunnya secara bertahap sesuai tingkat kematangannya. Namun, ada beberapa hal minimum yang sebaiknya sudah ada bahkan pada tahap awal, misalnya:

- kejelasan siapa yang menyetujui use case;
- kejelasan siapa pemilik data dan pemilik proses;
- dokumentasi dasar tentang tujuan, batas penggunaan, dan hasil pengujian;
- aturan dasar tentang penggunaan data sensitif dan vendor;
- serta mekanisme untuk melaporkan masalah atau insiden.

Seiring berkembangnya penggunaan AI, tata kelola dapat menjadi lebih matang, misalnya dengan:

- forum peninjauan lintas fungsi;
- risk register khusus AI;
- katalog use case AI organisasi;
- pola persetujuan berbasis tingkat risiko;
- kebijakan internal yang lebih rinci;
- dan mekanisme audit atau review berkala.

Pendekatan bertahap ini penting agar tata kelola tidak terasa terlalu berat di awal, tetapi juga tidak dibiarkan kosong.

1. Tata kelola AI dan kepemimpinan organisasi

Tata kelola AI juga sangat terkait dengan **kepemimpinan**. AI bukan hanya persoalan teknis, tetapi keputusan organisasi tentang bagaimana teknologi digunakan dalam kerja, layanan, dan pengambilan keputusan. Karena itu, pimpinan perlu memahami cukup baik:

- apa manfaat AI;
- apa risikonya;
- apa batasnya;
- dan bagaimana organisasi harus mengaturnya.

Pimpinan tidak harus menjadi ahli model atau arsitektur teknis, tetapi perlu mampu menilai:

- apakah arah penggunaan AI selaras dengan prioritas organisasi;
- apakah risiko tertentu layak diterima;
- apakah struktur tanggung jawab cukup jelas;
- dan apakah ada kapasitas yang cukup untuk mengelola sistem secara berkelanjutan.

Tanpa dukungan kepemimpinan, tata kelola AI cenderung lemah, terfragmentasi, atau terlalu bergantung pada inisiatif unit tertentu.

2. Tata kelola AI perlu membedakan jenis penggunaan

Tidak semua penggunaan AI memerlukan tata kelola yang identik. Organisasi perlu membedakan tingkat pengaturan berdasarkan:

- jenis use case;
- sensitivitas data;
- tingkat pengaruh terhadap proses atau layanan;
- dan sejauh mana AI berperan dalam workflow.

Sebagai contoh:

- penggunaan AI untuk ringkasan bahan internal mungkin memerlukan tata kelola yang lebih ringan;
- retrieval dokumen internal yang terbatas pada sumber resmi mungkin memerlukan pengaturan akses dan dokumentasi yang cukup jelas;
- penggunaan agentic AI yang terhubung ke beberapa tool atau vendor kemungkinan memerlukan tata kelola yang lebih ketat;
- dan use case yang menyentuh data pribadi atau dokumen sensitif perlu kontrol yang lebih kuat lagi.

Pendekatan seperti ini membantu organisasi tidak menerapkan satu aturan yang terlalu longgar untuk semua hal, tetapi juga tidak membuat semua use case menjadi terlalu berat tanpa alasan.

3. Tata kelola AI dan budaya akuntabilitas

Pada akhirnya, tata kelola AI akan lemah bila tidak didukung oleh **budaya akuntabilitas**. Organisasi perlu membangun kebiasaan bahwa penggunaan AI:

- harus bisa dijelaskan;
- harus bisa dipertanyakan;
- harus bisa diperbaiki;
- dan tidak boleh dibiarkan tumbuh di wilayah abu-abu tanpa pemilik yang jelas.

Budaya seperti ini penting karena AI mudah menimbulkan ilusi objektivitas. Hasil AI kadang terlihat netral dan meyakinkan, padahal tetap mengandung pilihan desain, keterbatasan data, dan risiko penggunaan. Tata kelola yang sehat menjaga agar organisasi tidak menerima AI secara pasif, tetapi tetap aktif menilai dan mengarahkannya.

4. Tata kelola AI sebagai pondasi keberlanjutan

Tata kelola juga penting untuk **keberlanjutan**. Banyak inisiatif AI gagal bukan pada saat awal, tetapi beberapa waktu setelah diterapkan, karena:

- tidak ada yang jelas memelihara sistem;
- perubahan vendor tidak dikelola;
- prompt atau model berubah tanpa review;
- pengguna baru tidak mendapat panduan yang cukup;
- atau insiden kecil diabaikan sampai akhirnya menumpuk.

Dengan tata kelola yang baik, organisasi lebih siap menjaga AI sebagai bagian dari sistem kerja jangka menengah dan panjang, bukan hanya proyek sesaat.

Tata kelola AI dalam organisasi pemerintah adalah kerangka yang memastikan bahwa AI digunakan secara terarah, sah, aman, etis, dan akuntabel. Tata kelola ini mencakup kebijakan dasar, pembagian peran, mekanisme penilaian dan persetujuan, dokumentasi, monitoring, evaluasi, dan keputusan tindak lanjut. Dalam sektor publik, tata kelola AI sangat penting karena penggunaan AI menyentuh layanan, data, proses, dan kepercayaan publik.

5.11 Kebijakan internal dan tata kelola vendor

Setelah membahas tata kelola AI dalam organisasi pemerintah secara lebih umum, langkah berikutnya adalah menerjemahkan tata kelola tersebut ke dalam dua wilayah yang sangat operasional, yaitu **kebijakan internal** dan **tata kelola vendor**. Keduanya penting karena banyak organisasi tidak gagal pada level prinsip, melainkan pada level pelaksanaan sehari-hari. Organisasi mungkin sudah sepakat bahwa AI harus aman, etis, akuntabel, dan diawasi manusia, tetapi tanpa kebijakan internal yang cukup jelas, para pengguna akan tetap kebingungan: apa yang boleh dimasukkan ke sistem, kapan AI boleh dipakai, kapan hasil AI harus diperiksa, siapa yang harus dimintai persetujuan, dan apa yang harus dilakukan bila hasil AI bermasalah. Demikian pula, tanpa tata kelola vendor yang memadai, organisasi dapat memakai layanan AI pihak ketiga secara cepat, tetapi kehilangan kendali atas data, perubahan model, atau penanganan insiden.

Dalam konteks sektor publik, dua hal ini sangat penting. AI jarang hanya menjadi urusan satu orang atau satu eksperimen kecil. Begitu ia masuk ke proses organisasi, penggunaan AI akan bersinggungan dengan:

- data dan dokumen internal;
- proses kerja resmi;
- aplikasi dan infrastruktur organisasi;
- pihak ketiga yang menyediakan model atau layanan;
- serta kebiasaan kerja pegawai yang perlu diarahkan agar tetap berada dalam jalur yang sehat.

Karena itu, tata kelola AI perlu diwujudkan dalam bentuk aturan dan pengaturan yang dapat dipakai sehari-hari. Di sinilah kebijakan internal dan tata kelola vendor memainkan peran yang sangat besar.

Penting untuk dipahami bahwa kebijakan internal tidak selalu harus langsung berbentuk dokumen yang sangat panjang dan sangat teknis. Yang lebih penting adalah bahwa kebijakan tersebut:

- cukup jelas;
- bisa dipahami oleh pengguna;
- relevan dengan use case yang memang muncul di organisasi;
- dan dapat dipakai sebagai acuan yang nyata.

Pada tahap awal, kebijakan internal bisa dimulai dari aturan dasar yang cukup sederhana, misalnya:

- jenis data atau dokumen yang tidak boleh dimasukkan ke AI tertentu;
- kewajiban memeriksa hasil AI sebelum dipakai lebih lanjut;
- kewajiban menggunakan sumber resmi untuk use case tertentu;
- larangan memakai AI sebagai penentu final pada konteks yang sensitif;
- dan kewajiban mencatat atau melaporkan penggunaan AI tertentu yang bersifat baru atau berisiko lebih tinggi.

Seiring organisasi semakin matang, kebijakan dapat diperluas menjadi lebih rinci, misalnya memuat klasifikasi use case, persyaratan pengujian, pengaturan vendor, pengelolaan prompt, pengelolaan logging, sampai prosedur evaluasi berkala.

Agar lebih operasional, organisasi dapat memikirkan beberapa unsur minimum yang sebaiknya ada dalam kebijakan internal AI.

1. Tujuan dan ruang lingkup

Kebijakan perlu menjelaskan mengapa organisasi menggunakan AI, untuk konteks apa, dan pada area mana kebijakan ini berlaku. Ini membantu pengguna memahami bahwa AI dipakai sebagai alat bantu yang berada dalam kerangka tujuan organisasi, bukan sekadar alat bebas pakai.

2. Prinsip dasar penggunaan

Kebijakan sebaiknya menegaskan prinsip seperti:

- kebermanfaatan;
- akuntabilitas;
- perlindungan data;
- pengawasan manusia;
- penggunaan sumber yang sah;
- dan pembatasan penggunaan pada konteks yang tidak layak.

Bagian ini menghubungkan kebijakan operasional dengan nilai dan tata kelola yang lebih luas.

3. Aturan mengenai data dan dokumen

Kebijakan perlu menjelaskan:

- jenis data apa yang boleh dan tidak boleh diproses dengan AI tertentu;
- bagaimana data sensitif diperlakukan;
- kapan de-identification atau masking diperlukan;
- dan apa syarat penggunaan dokumen internal, data pribadi, atau informasi yang belum final.

Bagian ini sangat penting karena banyak risiko AI justru muncul dari cara pengguna memasukkan data ke sistem.

4. Aturan penggunaan hasil AI

Kebijakan juga perlu menjelaskan:

- kapan hasil AI hanya boleh diperlakukan sebagai bantuan awal;
- kapan hasil harus diverifikasi;
- kapan hasil tidak boleh dipakai tanpa persetujuan manusia;
- dan apakah pengguna wajib menelusuri sumber asli pada jenis tugas tertentu.

Ini penting agar AI tidak diperlakukan sebagai otoritas final.

5. Peran dan tanggung jawab

Kebijakan perlu menjelaskan siapa:

- pengguna;
- pemilik proses;
- pengelola data;
- tim teknis;
- pihak yang menyetujui use case;
- dan pihak yang menangani risiko atau insiden.

Tanpa bagian ini, kebijakan akan sulit dijalankan secara konsisten.

6. Pelaporan masalah dan insiden

Organisasi perlu menetapkan bagaimana pengguna melaporkan:

- keluaran yang salah;
- potensi kebocoran data;
- penggunaan AI di luar tujuan;
- kegagalan guardrail;
- atau masalah lain yang berkaitan dengan AI.

Ini penting agar organisasi bisa belajar dan memperbaiki sistem.

7. Penggunaan vendor dan layanan eksternal

Karena banyak AI modern bergantung pada pihak ketiga, kebijakan internal juga perlu menjelaskan:

- kapan layanan eksternal boleh dipakai;
- siapa yang menyetujuinya;
- jenis data apa yang tidak boleh dikirim ke vendor;
- dan apa proses penilaiannya.

Dengan memasukkan unsur ini, kebijakan internal tidak berhenti pada perilaku pengguna, tetapi juga menghubungkan pengguna dengan tata kelola vendor.

5.11.1 Prinsip tata kelola vendor AI

Agar pengelolaan vendor lebih sehat, organisasi dapat memakai beberapa prinsip dasar.

1. Due diligence sebelum penggunaan

Sebelum memakai vendor, organisasi perlu melakukan penilaian yang cukup terhadap:

- keamanan;
- privasi;
- pengelolaan data;
- perubahan model;
- logging;
- dukungan insiden;
- dan kesesuaian dengan kebutuhan tata kelola organisasi.

Ini sejalan dengan penekanan **NIST AI 600-1** tentang *due diligence* terhadap privasi data, keamanan, risiko pihak ketiga, inventarisasi pihak ketiga yang memiliki akses ke konten organisasi, serta pencatatan perubahan dan rencana kontinjensi.

2. Batas data yang boleh dibagikan

Organisasi perlu menetapkan secara jelas data apa yang boleh dan tidak boleh dikirim ke vendor. Ini harus terhubung dengan kebijakan internal, bukan hanya berdasarkan kenyamanan pengguna.

3. Kejelasan kontraktual dan operasional

Hubungan dengan vendor perlu menjelaskan:

- tanggung jawab tiap pihak;
- penanganan insiden;
- perubahan layanan;
- hak audit atau visibilitas tertentu;
- dan kewajiban vendor terhadap keamanan serta perlindungan data.

4. Monitoring vendor secara berkelanjutan

Tata kelola vendor tidak berhenti saat kontrak ditandatangani. Organisasi perlu terus memantau:

- perubahan model atau layanan;
- masalah performa;
- insiden keamanan;
- dan dampak perubahan tersebut terhadap use case organisasi.

5. Contingency dan fallback

Organisasi perlu punya rencana cadangan bila vendor:

- tidak tersedia,
- mengubah layanan secara signifikan,
- atau menimbulkan risiko yang tidak lagi dapat diterima.

Tanpa fallback, organisasi terlalu rapuh terhadap perubahan pihak ketiga.

Kebijakan internal dan tata kelola vendor tidak boleh berjalan sendiri-sendiri. Keduanya harus saling terhubung. Misalnya:

- jika kebijakan internal melarang data sensitif dimasukkan ke layanan eksternal, maka tata kelola vendor harus memastikan aturan itu dapat diterapkan;
- jika vendor mengubah model, organisasi perlu mekanisme internal untuk menilai dampaknya pada workflow;
- jika terjadi insiden pada layanan vendor, pengguna internal harus tahu ke mana melapor dan bagaimana menyesuaikan penggunaan AI untuk sementara.

Dengan kata lain, hubungan antara pengguna, pemilik proses, tim teknis, dan vendor harus dibingkai oleh aturan yang cukup jelas dan saling mendukung.

Satu prinsip penting yang perlu ditekankan adalah bahwa penggunaan vendor **tidak memindahkan tanggung jawab organisasi**. Organisasi tetap bertanggung jawab atas:

- keputusan memakai AI;
- konteks penggunaan;
- data yang dibagikan;
- batas penggunaan hasil;
- dan dampak AI pada proses kerja serta layanan.

Vendor bisa membantu menyediakan teknologi, tetapi tidak menggantikan kewajiban organisasi untuk menjaga akuntabilitas, keamanan, etika, dan kepercayaan publik.

5.11.2 Kebijakan dan tata kelola seiring kematangan organisasi

Pada tahap awal, organisasi mungkin hanya memerlukan aturan dasar dan proses vendor yang sederhana. Namun, seiring penggunaan AI berkembang, kebijakan internal dan tata kelola vendor perlu ikut berkembang. Organisasi bisa mulai dari:

- aturan dasar penggunaan data;
- kewajiban verifikasi hasil AI;
- dan persetujuan dasar untuk vendor tertentu.

Seiring waktu, organisasi mungkin perlu menambah:

- klasifikasi use case berdasarkan tingkat risiko;
- daftar vendor yang disetujui;
- evaluasi berkala vendor;
- review perubahan model;
- proses penilaian khusus untuk agentic AI;
- dan mekanisme audit atau penjaminan mutu yang lebih kuat.

Pendekatan bertahap seperti ini penting agar organisasi tidak menunggu sampai semua hal sempurna untuk mulai bergerak, tetapi juga tidak membiarkan AI tumbuh tanpa aturan yang cukup.

Kebijakan internal dan tata kelola vendor adalah dua pilar operasional yang sangat penting dalam tata kelola AI. Kebijakan internal membantu mengarahkan bagaimana AI digunakan sehari-hari oleh pegawai, data apa yang boleh diproses, bagaimana hasil AI diposisikan, dan bagaimana insiden dilaporkan. Tata kelola vendor membantu organisasi menjaga kendali ketika AI melibatkan pihak ketiga, terutama dalam hal data, perubahan layanan, risiko keamanan, dan akuntabilitas.

5.12 Rangkuman

Bab ini menempatkan **keamanan, etika, risiko, dan tata kelola AI** sebagai pagar utama bagi seluruh penggunaan AI di organisasi pemerintah. Jika bab-bab sebelumnya membahas apa itu AI, bagaimana memilih use case yang layak, dan bagaimana merancang serta mengintegrasikan solusi AI, maka Bab 5 menegaskan bahwa semua itu tidak akan cukup tanpa kerangka pengamanan, pengawasan, dan tata kelola yang memadai. Dalam sektor publik, AI tidak boleh diperlakukan sekadar sebagai inovasi teknis, melainkan sebagai kemampuan organisasi yang harus berada di dalam batas nilai publik, akuntabilitas, keamanan, dan perlindungan terhadap warga serta institusi. Hal ini sejalan dengan prinsip pelaksanaan AI nasional yang menekankan kebermanfaatan, keadilan, akuntabilitas, perlindungan data pribadi, transparansi, keamanan, integritas, inklusivitas, dan keterlibatan manusia.

Pada bagian awal, bab ini menjelaskan bahwa **keamanan AI berbeda dari keamanan sistem biasa**. Solusi AI memang mewarisi kebutuhan keamanan TI yang lazim, seperti kontrol akses, keamanan aplikasi, dan perlindungan jaringan. Namun, AI juga membawa lapisan risiko tambahan karena ia memiliki aset yang lebih beragam, seperti model, prompt, corpus dokumen, log interaksi, dan alur inferensi. AI juga cenderung kurang sepenuhnya dapat diprediksi dibanding sistem biasa, memiliki permukaan serangan yang lebih luas, serta dapat dipengaruhi bukan hanya oleh data, tetapi juga oleh instruksi dalam bentuk prompt. Pedoman pengembangan AI yang aman menegaskan bahwa sistem AI menghadapi kerentanan baru di samping ancaman keamanan siber yang sudah lazim, dan bahwa keamanan harus diterapkan sepanjang siklus hidup AI.

Bab ini kemudian memetakan **risiko pada data, model, output, prompt, dan vendor**. Risiko pada data meliputi kualitas yang buruk, bias, pencemaran data, penggunaan data sensitif, dan ketidakjelasan asal-usul data. Risiko pada model meliputi ketidakcocokan dengan use case, perubahan perilaku akibat pembaruan, kerapuhan terhadap variasi input, serta keterbatasan kemampuan penjelasan. Risiko pada output meliputi keluaran yang meyakinkan tetapi keliru, keluaran yang melampaui ruang lingkup, kebocoran informasi, dan pengaruh berlebihan terhadap tindakan manusia. Risiko pada prompt berkaitan dengan prompt yang terlalu umum, prompt injection, prompt yang tidak terdokumentasi, dan ketidakkonsistenan penggunaan. Sementara itu, risiko pada vendor mencakup ketergantungan, kurangnya visibilitas, persoalan data, perubahan layanan, dan supply chain risk. Seluruh titik risiko ini perlu dibaca secara terhubung, bukan terpisah-pisah.

Setelah itu, pembahasan bergerak ke **privasi dan perlindungan data**. Pada bagian ini dijelaskan bahwa AI membuat isu privasi menjadi lebih sensitif karena bekerja dengan volume data yang besar, mampu membentuk inferensi dari kombinasi data, dan dapat menampilkan kembali informasi sensitif melalui keluaran bahasa alami. Karena itu, organisasi perlu menerapkan prinsip bahwa hanya data yang benar-benar diperlukan yang dipakai, akses dibatasi sesuai kebutuhan, data sensitif diperlakukan secara khusus, dan bila perlu dilakukan de-identification, masking, atau redaction. Pembahasan juga menekankan bahwa perlindungan data dalam AI bukan hanya persoalan penyimpanan data, tetapi juga persoalan desain output, kontrol penggunaan vendor, dokumentasi sumber data, dan pengawasan manusia terhadap apa yang dimasukkan maupun dikeluarkan sistem. Prinsip

perlindungan data pribadi sebagai bagian dari pelaksanaan AI nasional juga menjadi dasar penting di sini.

Bab ini kemudian membahas **bias, fairness, dan integritas informasi**. Bias dijelaskan sebagai kecenderungan sistem menghasilkan pola yang tidak seimbang atau tidak semestinya terhadap kelompok, jenis kasus, atau konteks tertentu. Fairness dipahami sebagai upaya menjaga agar sistem tidak menghasilkan perlakuan yang tidak adil atau tidak proporsional. Sementara itu, integritas informasi berkaitan dengan kemampuan organisasi menjaga agar AI tidak merusak mutu informasi yang dipakai, diringkas, atau dihasilkan. Dalam konteks sektor publik, ketiga isu ini sangat penting karena AI dapat mempengaruhi cara dokumen dibaca, aduan diklasifikasikan, isu kebijakan diringkas, dan perhatian organisasi diarahkan. Karena itu, organisasi perlu menguji variasi kasus, mewajibkan sumber resmi pada use case tertentu, menangkap koreksi pengguna, dan memantau perubahan kualitas hasil dari waktu ke waktu.

Pada bagian **risiko khusus generative AI dan agentic AI**, bab ini menyoroti lima risiko yang perlu dibaca secara khusus, yaitu:

- **confabulation,**
- **prompt injection,**
- **data leakage,**
- **over-automation,**
- **dan tindakan agent yang tidak terkendali.**

Confabulation merujuk pada keluaran yang tampak masuk akal tetapi tidak didukung sumber yang cukup benar. Prompt injection berkaitan dengan upaya memanipulasi perilaku model melalui instruksi yang tidak semestinya. Data leakage dapat terjadi melalui input, output, log, atau integrasi dengan vendor dan tool. Over-automation menggambarkan kecenderungan menyerahkan terlalu banyak langkah kerja kepada AI hanya karena AI tampak mampu. Sementara itu, tindakan agent yang tidak terkendali menggambarkan bahaya ketika AI multi-langkah diberi ruang gerak terlalu luas tanpa batas kewenangan yang tegas. Pada bagian ini ditekankan bahwa organisasi perlu membangun guardrail, membatasi ruang gerak agent, dan menetapkan titik pemeriksaan manusia secara jelas.

Bab ini juga menempatkan **prinsip etika AI dan nilai publik** sebagai landasan normatif. Prinsip-prinsip seperti kebermanfaatan, keadilan, penghormatan terhadap martabat manusia, transparansi, akuntabilitas, perlindungan data pribadi, keamanan, keselamatan, integritas, inklusivitas, serta keterlibatan dan pengawasan manusia dibahas sebagai dasar yang perlu membimbing seluruh penggunaan AI. Dalam konteks sektor publik, nilai publik menjadi pusat dari seluruh prinsip tersebut. AI yang baik bukan hanya AI yang bekerja secara teknis, tetapi AI yang memperkuat layanan, akses informasi, tata kelola, dan kemampuan organisasi tanpa mengorbankan hak, kepercayaan, dan kualitas keputusan. Konsep Pedoman Etika Kecerdasan Artifisial menegaskan bahwa etika AI mencakup nilai etika, manfaat penerapan etika, identifikasi dan mitigasi risiko, langkah-langkah perlindungan, peran dan tanggung jawab para pihak, serta pemantauan dan evaluasi.

Bagian berikutnya membahas **human oversight dan safeguard**. Di sini dijelaskan bahwa keterlibatan manusia perlu dirancang secara nyata, bukan sekadar diasumsikan. Manusia perlu hadir dalam penetapan tujuan, pemeriksaan hasil, koreksi keluaran, persetujuan

langkah berikutnya, dan pengambilan keputusan akhir bila diperlukan. Sementara itu, safeguard mencakup pembatasan akses, guardrails prompt, pembatasan output, pembatasan ruang gerak agent, logging, fallback, dan mekanisme kontrol lain yang dirancang untuk menjaga AI tetap berada dalam ruang penggunaan yang sah. Prinsip keterlibatan dan pengawasan manusia yang ditekankan dalam kebijakan nasional menjadi dasar penting di sini.

Selanjutnya, bab ini menjelaskan **manajemen risiko AI** sebagai proses yang sistematis dan berkelanjutan untuk mengidentifikasi, menilai, mengendalikan, memantau, dan meninjau ulang risiko AI sepanjang siklus hidup sistem. Organisasi diajak melihat bahwa risiko AI perlu dikelola sejak tahap pemilihan use case, pemilihan data, desain arsitektur, pemilihan vendor, integrasi ke workflow, pengujian, sampai operasi harian dan perubahan sistem. Pada bagian ini juga dibahas pentingnya risk register, penetapan pemilik risiko, pengelolaan risiko residual, serta budaya organisasi yang cukup matang untuk menyeimbangkan inovasi dan kontrol.

Bab ini kemudian membahas **tata kelola AI dalam organisasi pemerintah**. Tata kelola dipahami sebagai struktur keputusan dan aturan main yang memastikan AI digunakan secara sah, terarah, aman, etis, dan akuntabel. Tata kelola ini mencakup kebijakan dan prinsip penggunaan, pembagian peran dan tanggung jawab, mekanisme penilaian dan persetujuan, dokumentasi, monitoring, evaluasi, dan tindak lanjut. Pembahasan ini menegaskan bahwa tata kelola AI bukan hanya urusan tim teknis, melainkan persoalan lintas fungsi yang melibatkan pemilik proses, pengelola data, keamanan informasi, pengawasan, pimpinan, dan pengguna sistem.

Terakhir, bab ini membahas **kebijakan internal dan tata kelola vendor**. Kebijakan internal diperlukan agar penggunaan AI sehari-hari tidak berkembang secara liar berdasarkan kebiasaan individual, tetapi diarahkan melalui aturan yang cukup jelas tentang data, penggunaan hasil AI, kewajiban verifikasi, pelaporan masalah, dan ruang lingkup penggunaan. Sementara itu, tata kelola vendor penting karena banyak organisasi memakai model, API, atau layanan AI pihak ketiga. Organisasi perlu melakukan due diligence, membatasi data yang dibagikan, menetapkan kejelasan kontraktual dan operasional, memantau vendor secara berkelanjutan, dan menyiapkan fallback. **NIST AI 600-1** menekankan inventarisasi pihak ketiga yang memiliki akses ke konten organisasi, evaluasi risiko pemasok, pencatatan perubahan, due diligence atas privasi dan keamanan, serta rencana kontinjensi bila layanan pihak ketiga mengalami masalah.

Secara keseluruhan, Bab 5 menegaskan bahwa keamanan, etika, risiko, dan tata kelola AI bukan pelengkap di tepi penggunaan AI, tetapi pondasi yang harus menyertai seluruh pemanfaatan AI di organisasi pemerintah. AI yang baik bukan hanya AI yang berfungsi, tetapi AI yang aman, adil, diawasi, terdokumentasi, dan dipakai dalam kerangka yang melindungi organisasi, pengguna, dan masyarakat.

5.13 Pertanyaan refleksi

Jawablah pertanyaan-pertanyaan berikut berdasarkan konteks instansi, unit kerja, atau pengalaman Anda sendiri.

1. Mengapa keamanan AI tidak dapat diperlakukan sama dengan keamanan sistem informasi biasa?
2. Dari lima titik risiko yang dibahas—data, model, output, prompt, dan vendor—mana yang menurut Anda paling rentan di organisasi Anda saat ini? Mengapa?
3. Apakah unit kerja Anda sudah memiliki kebiasaan yang cukup aman dalam menggunakan AI terhadap data internal atau data pribadi? Jelaskan.
4. Dalam konteks pekerjaan Anda, bagaimana bias AI bisa muncul walaupun sistem tampak bekerja dengan baik?
5. Mengapa integritas informasi perlu menjadi perhatian utama ketika AI dipakai untuk ringkasan, tanya jawab berbasis dokumen, atau drafting awal?
6. Dari risiko khusus generative AI dan agentic AI—confabulation, prompt injection, data leakage, over-automation, dan tindakan agent yang tidak terkendali—mana yang paling mungkin terjadi di lingkungan kerja Anda?
7. Menurut Anda, apa perbedaan antara keberadaan manusia di dalam proses dan *human oversight* yang benar-benar efektif?
8. Apakah organisasi Anda sudah memiliki bentuk tata kelola AI, meskipun mungkin belum disebut dengan istilah itu? Jelaskan bentuknya.
9. Mengapa kebijakan internal AI perlu dibuat, walaupun organisasi baru menggunakan AI pada use case yang tampak sederhana?
10. Bila organisasi Anda memakai vendor atau layanan AI pihak ketiga, informasi apa saja yang menurut Anda harus diperiksa sebelum layanan itu dipakai?

5.14 Latihan dan tugas kecil

Latihan 1 — Memetakan risiko AI pada satu use case

Pilih satu use case AI yang menurut Anda relevan untuk unit kerja Anda. Lalu petakan secara singkat risikonya pada lima titik berikut:

1. risiko pada data;
2. risiko pada model atau engine;
3. risiko pada output;
4. risiko pada prompt atau instruksi;
5. risiko pada vendor atau pihak ketiga.

Setelah itu, tentukan:

- risiko mana yang paling besar;
- dan kontrol awal apa yang menurut Anda paling perlu diterapkan.

Tujuan latihan ini adalah membantu Anda berpindah dari pandangan umum tentang risiko AI menuju pemetaan yang lebih konkret dan operasional.

Latihan 2 — Menilai privasi dan perlindungan data

Ambil satu contoh use case AI, lalu jawab pertanyaan berikut:

1. data apa saja yang akan diproses;
2. apakah data tersebut mengandung data pribadi atau data sensitif;
3. apakah semua data itu benar-benar diperlukan;
4. apakah ada bagian yang bisa dihapus, disamarkan, atau dibatasi;
5. siapa yang seharusnya boleh mengakses data dan output sistem;
6. apakah ada vendor eksternal yang akan menerima data tersebut.

Di akhir latihan, simpulkan:

- apakah use case itu cukup aman dari sisi privasi;
- atau perlu desain ulang sebelum dijalankan.

Tujuan latihan ini adalah membantu Anda melihat perlindungan data sebagai bagian dari desain AI, bukan sekadar masalah administratif.

Latihan 3 — Menilai fairness dan integritas informasi

Pilih satu use case AI yang menghasilkan klasifikasi, ringkasan, retrieval, atau drafting. Lalu uraikan:

1. di titik mana bias bisa muncul;
2. kelompok, jenis kasus, atau jenis dokumen apa yang mungkin lebih rentan dirugikan;
3. bagaimana AI dapat merusak integritas informasi pada use case tersebut;
4. apa yang dapat dilakukan manusia untuk memeriksa dan mengoreksi hasilnya.

Tujuan latihan ini adalah membantu Anda memahami bahwa AI yang tampak bekerja baik pun tetap bisa membawa risiko ketidakadilan dan penurunan mutu informasi.

Latihan 4 — Mendesain human oversight dan safeguard

Pilih satu use case AI, lalu buat rancangan singkat:

1. siapa yang menetapkan tujuan penggunaan AI;
2. siapa yang memeriksa hasilnya;
3. kapan hasil AI boleh dipakai langsung sebagai bantuan awal;
4. kapan hasil AI harus diverifikasi lebih ketat;
5. safeguard apa yang perlu ada pada data, prompt, output, atau tindakan agent;
6. apa fallback jika AI gagal atau hasilnya diragukan.

Tujuan latihan ini adalah membantu Anda menerjemahkan prinsip etika dan pengawasan manusia ke dalam desain kerja yang nyata.

Latihan 5 — Menyusun kerangka tata kelola sederhana

Bayangkan organisasi Anda ingin mulai memakai AI secara lebih sistematis. Buat kerangka tata kelola sederhana yang memuat:

1. siapa yang boleh mengusulkan use case AI;
2. siapa yang menilai manfaat dan risiko;
3. siapa yang menyetujui penggunaan;
4. dokumen atau informasi apa yang wajib dicatat;
5. bagaimana insiden dilaporkan;
6. kapan sistem harus ditinjau ulang.

Tujuan latihan ini adalah membantu Anda melihat bahwa tata kelola AI dapat mulai dibangun secara bertahap, tidak harus menunggu organisasi menjadi sangat matang.

Tugas Kecil

Judul: Menyusun Kerangka Pengamanan dan Tata Kelola AI untuk Unit Kerja Saya

Tuliskan esai singkat sepanjang **2–3 halaman** yang memuat unsur-unsur berikut:

1. **Use case AI yang dipilih**

Jelaskan satu use case AI yang menurut Anda paling realistis atau paling dekat dengan pekerjaan unit Anda.

2. **Profil risiko awal**

Uraikan risiko utamanya dari sisi:

- data,
- model,
- output,
- prompt,
- dan vendor.

3. **Privasi, fairness, dan integritas informasi**

Jelaskan:

- potensi masalah privasi;
- potensi bias atau ketidakadilan;
- dan potensi gangguan pada integritas informasi.

4. **Risiko khusus generative AI atau agentic AI**

Bila relevan, jelaskan apakah use case ini berpotensi terkena:

- confabulation,
- prompt injection,
- data leakage,
- over-automation,
- atau tindakan agent yang tidak terkendali.

5. **Human oversight dan safeguard**

Tuliskan siapa yang memeriksa hasil AI, apa titik persetujuannya, dan safeguard apa yang perlu diterapkan pada data, prompt, output, atau workflow.

6. **Kerangka manajemen risiko**

Jelaskan bagaimana risiko akan:

- diidentifikasi,
- dinilai,
- dikendalikan,
- dipantau,
- dan ditinjau ulang.

7. **Tata kelola organisasi**

Jelaskan siapa yang perlu dilibatkan dalam pengelolaan use case ini, bagaimana pembagian tanggung jawabnya, dan dokumen apa yang perlu tersedia.

8. **Kebijakan internal dan vendor**

Jelaskan aturan internal minimum apa yang perlu ada, dan bila vendor terlibat, apa yang harus diperiksa serta dibatasi.

9. **Kesimpulan**

Jelaskan apakah menurut Anda use case ini:

- cukup aman untuk dimulai,
- perlu dipersempit,

- perlu kontrol tambahan,
- atau sebaiknya ditunda sampai organisasi lebih siap.

Tujuan tugas kecil

Tugas ini bertujuan membantu Anda:

- menghubungkan konsep keamanan, etika, risiko, dan tata kelola AI dengan kondisi nyata organisasi;
- melihat bahwa AI yang layak tidak hanya berguna, tetapi juga harus dapat dikendalikan;
- dan menyiapkan dasar berpikir untuk bab berikutnya, yang dapat bergerak ke evaluasi, implementasi bertahap, pengukuran manfaat, atau pengembangan kapasitas organisasi dalam adopsi AI.