Etika dan Keamanan Kecerdasan Buatan

1. Bias pada Data

Model kecerdasan buatan mewarisi bias yang ada pada data latihnya. Contoh yang sering dikutip adalah sistem seleksi pelamar kerja milik Amazon yang dihentikan pada tahun 2018 karena cenderung merendahkan pelamar perempuan, akibat dilatih dengan data riwayat perekrutan yang didominasi laki-laki.

Mitigasi bias mencakup audit dataset, pengukuran metrik kewajaran antar kelompok, serta pengujian model pada subkelompok yang berbeda sebelum digunakan.

2. Privasi dan Pelindungan Data Pribadi

Di Indonesia, pelindungan data pribadi diatur dalam Undang-Undang Nomor 27 Tahun 2022 tentang Pelindungan Data Pribadi. Prinsip pentingnya adalah minimisasi data, persetujuan yang eksplisit dari pemilik data, serta hak subjek data untuk mengakses dan meminta penghapusan datanya.

Anonimisasi menghapus identitas secara permanen, sedangkan pseudonimisasi hanya mengganti identitas dengan kode yang masih dapat ditelusuri kembali. Keduanya mengurangi risiko, tetapi hanya anonimisasi yang benar-benar melepaskan data dari pemiliknya.

3. Transparansi dan Keterjelasan

Model kotak hitam sulit dijelaskan kepada pengguna yang terdampak keputusannya. Teknik seperti LIME dan SHAP memperkirakan seberapa besar kontribusi tiap fitur terhadap satu prediksi. Keterjelasan penting pada bidang berisiko tinggi seperti kredit, kesehatan, dan hukum.

4. Serangan Adversarial

Serangan adversarial mengubah input sedikit saja, sering kali dengan derau yang tidak terlihat oleh mata, sehingga model salah mengklasifikasi. Contoh terkenal adalah stiker kecil pada rambu stop yang membuat model penglihatan komputer membacanya sebagai rambu batas kecepatan. Pertahanan yang umum adalah adversarial training, yaitu melatih model dengan contoh serangan.

5. Prompt Injection

Prompt injection adalah serangan berupa instruksi berbahaya yang disisipkan dalam masukan pengguna atau dalam dokumen, dengan tujuan membajak perilaku model. Risikonya besar pada sistem RAG karena dokumen yang diambil diperlakukan sebagai konteks tepercaya.

Mitigasinya antara lain memisahkan instruksi sistem dari data, menyaring keluaran model, dan menerapkan hak akses minimum pada alat yang dapat dipanggil model.

Tabel 1. Ringkasan risiko dan mitigasi

Risiko	Contoh	Mitigasi
Bias	Seleksi otomatis menolak kelompok tertentu	Audit dataset
Kebocoran data	Model mengungkap data pribadi	Anonimisasi
Serangan adversarial	Stiker pada rambu stop	Adversarial training
Prompt injection	Instruksi tersembunyi di dokumen	Pisahkan instruksi dari data

Sumber: ringkasan penulis.

6. Tata Kelola dan Akuntabilitas

Tata kelola yang baik mencakup audit berkala, keterlibatan manusia dalam pengambilan keputusan penting (human in the loop), dan pencatatan log keputusan. Dokumentasi model dalam bentuk model card diperkenalkan oleh Mitchell dan kawan-kawan pada tahun 2019 untuk menjelaskan tujuan, data, dan batasan sebuah model.

Akuntabilitas tetap berada pada manusia dan organisasi yang menerapkan sistem, bukan pada model itu sendiri.