BIAS OTOMASI: SEKADAR TUKANG STEMPEL

"…the more advanced a control system is, so the more crucial may be the contribution of the human operator."

"Semakin canggih sebuah sistem kendali, bisa jadi semakin penting pula peran operator manusianya."

— Lisanne Bainbridge, psikolog, 1983[101]

> **Pertanyaan bab ini:** Bila AI hampir selalu benar dan seorang manusia ditugasi memeriksanya, apakah pemeriksaan itu benar-benar terjadi?
>
> **Jawaban singkat:** Sering kali tidak. Pemeriksa yang jarang menemukan kesalahan pelan-pelan berhenti mencari, dan saran mesin yang keliru lolos, bahkan di tangan para ahli. Pemeriksaan manusia baru berarti bila ditempatkan di keputusan yang tepat dan dirancang supaya benar-benar terjadi.
>
> **Alat yang Anda bawa pulang:** empat jenis keputusan untuk menentukan di mana manusia memeriksa, tiga kontrol untuk proses pemberian persetujuan, dan satu eksperimen kecil untuk mengetahui apakah para pemeriksa di kantor Anda masih benar-benar memeriksa.

Identitas Dokumen

BUKU	**Stay Human. Stay in Charge.** Transformasi AI di Korporasi: Maksimalkan Hasil Tanpa Halusinasi dan Kebocoran Data
PENULIS	Dian Martin, Ketua Umum Asosiasi AI Indonesia
BAGIAN	Bagian Kedua: Human In Charge
BAB	Bab 6: Bias Otomasi: Sekadar Tukang Stempel
HALAMAN	131–160 (nomor halaman cetak buku)
ISI BERKAS	Isi bab, Panduan Daftar Pustaka, dan Daftar Pustaka bab ini

SELASA, 2 MEI 2023. Perhimpunan radiologi Amerika Utara, RSNA, mengumumkan hasil sebuah eksperimen yang terbit di jurnal mereka, *Radiology*.[102,103]

Para peneliti dari Jerman dan Belanda, dipimpin Thomas Dratsch dari Rumah Sakit Universitas Köln, mengundang 27 radiolog. Tugas mereka membaca 50 mamogram, foto rontgen payudara untuk skrining kanker. Untuk setiap foto, mereka menentukan kategori temuan menurut BI-RADS, sistem baku yang dipakai para radiolog untuk menggolongkan temuan pada citra payudara. Kategori itu bukan diagnosis, tetapi penting untuk menentukan langkah perawatan berikutnya.

Setiap foto disertai saran kategori dari sebuah sistem AI.

Sepuluh foto pertama adalah latihan, dan di sana saran AI selalu benar. Lalu datang 40 foto berikutnya. Pada 12 di antaranya, para peneliti diam-diam menyelipkan saran yang keliru, seolah-olah berasal dari AI.

Radiolog pemula menjawab benar hampir 80 persen ketika saran AI benar. Ketika sarannya keliru, jawaban benar mereka jatuh ke kurang dari 20 persen. Itu masih bisa dimaklumi; mereka sedang belajar.

Yang lebih mengusik adalah kelompok paling senior, yang rata-rata sudah berpraktik lebih dari 15 tahun. Ketika saran AI benar, ketepatan mereka 82 persen. Ketika saran AI keliru, ketepatannya turun menjadi 45,5 persen.[103] Dari setiap sepuluh mamogram yang disertai saran keliru, radiolog yang sudah belasan tahun membaca foto semacam itu menilai dengan benar kurang dari lima.

Dratsch sendiri memberi peringatan: karena skrining mamografi bersifat berulang dan sangat terstandar, bias otomasi bisa menjadi masalah ketika AI dimasukkan ke alur kerjanya.[103]

Bias otomasi (*automation bias*) adalah kecenderungan manusia untuk terlalu mengandalkan otomasi.[104] Saran mesin ditelan mentah-mentah, artinya diterima tanpa diperiksa. Penyebabnya jarang soal malas atau bodoh.

Di sebuah kantor distribusi, gejala yang sama muncul dalam bentuk yang jauh lebih biasa.

“Bulan ini nol lagi, Pak.”

Bu Rini meletakkan selembar rekap di meja Pak Hendra. Sudah tiga bulan asisten AI di PT Contoh Distribusi menjawab pertanyaan pelanggan soal tanggal jatuh tempo, sesuai lembar aturan pakai yang ditandatangani di pembahasan sebelumnya. Aturannya jelas: setiap minggu, contoh jawaban AI diperiksa manusia.

“Nol kesalahan?” Pak Hendra tersenyum. “Bagus, kan?”

“Itu yang membuat saya ragu.” Bu Rini menunjuk satu baris. “Uji kita dulu bilang AI ini benar 97 dari 100. Kalau setiap minggu yang diperiksa 50 jawaban, mestinya rata-rata ada satu atau dua yang keliru. Lima minggu berturut-turut pemeriksa kita tidak menemukan satu pun.”

Pak Hendra berhenti tersenyum. Ia mengambil rekap itu dan membacanya sekali lagi, kali ini pelan-pelan.

“Siapa yang memeriksa?”

“Staf penagihan kita yang paling senior, Pak. Orangnya teliti. Justru karena itu saya heran.”

“Jadi menurut Ibu, AI-nya yang mendadak sempurna, atau pemeriksanya yang berhenti melihat?”

Bu Rini tidak langsung menjawab. Dua-duanya mungkin. Yang jelas, rekap berisi angka nol itu tidak bisa membedakan keduanya.

Perusahaan, tokoh, dan angka dalam kisah PT Contoh Distribusi ini rekaan; bila ada kesamaan dengan perusahaan atau orang sungguhan, itu tidak disengaja. Bagian yang nyata, yaitu temuan para peneliti, saya sebut bersama sumbernya.

Banyak perusahaan menjawab setiap risiko AI dengan satu kalimat yang menenangkan: tetap ada manusia yang memeriksa. Apakah manusia itu benar-benar memeriksa, atau sekadar membubuhkan stempel?

Satpam dan Ribuan Tas: Batas Perhatian Manusia

Ambil contoh satpam di pintu masuk sebuah pusat perbelanjaan.

Pada hari pertama bertugas, ia memeriksa setiap tas dengan saksama. Ia membuka ritsleting, menyorotkan senter, dan memperhatikan isinya. Seminggu kemudian, ribuan tas sudah lewat dan tidak satu pun berisi barang berbahaya. Pelan-pelan, pemeriksaannya berubah. Senter hanya diarahkan sekilas. Lalu cukup melirik. Akhirnya tinggal melambaikan tangan.

Apakah satpam itu malas? Belum tentu. Ia sedang melakukan apa yang dilakukan otak manusia pada umumnya: berhenti memperhatikan sesuatu yang hampir selalu aman.

Lisanne Bainbridge, psikolog di University College London, sudah menulis soal ini pada 1983, jauh sebelum ada AI generatif. Dalam makalahnya tentang ironi otomasi, ia mencatat sebuah ironi yang sampai kini terasa segar. Perancang yang berusaha menyingkirkan operator manusia ternyata tetap menyisakan tugas untuk operator itu, dan salah satunya adalah memantau apakah sistem otomatisnya bekerja

dengan benar.[101] Mesin dipasang karena manusia dianggap kurang andal, lalu manusia yang sama diminta mengawasi mesin.

Ia juga mencatat batas kemampuan manusia untuk tugas semacam itu. Merujuk penelitian kewaspadaan sejak 1950, ia menulis bahwa bahkan orang yang sangat termotivasi pun tidak mungkin menjaga perhatian visual yang efektif lebih dari sekitar setengah jam, bila yang diawasinya jarang berubah.[101,105]

Setengah jam. Kini ganti satpam itu dengan karyawan yang setiap hari memeriksa ratusan jawaban AI, dan sebagian besar jawaban itu benar. Kewaspadaannya akan menurun seperti pemeriksaan si satpam.

Di dalam negeri, ada contoh dengan arah sebaliknya. Grup MIND ID, induk perusahaan tambang milik negara, menyebut telah memakai sistem *fatigue management* (pengelolaan kelelahan) berbasis AI yang memberi peringatan dini ketika mendeteksi indikasi kelelahan maupun *microsleep*, yaitu tertidur sesaat tanpa disadari, pada operator.[106] Menurut tafsir saya, di sini mesin dipasang untuk menjaga perhatian manusia, karena perusahaan tambang tahu bahwa perhatian operatornya punya batas. Pertanyaannya, apakah kantor yang meminta karyawannya memeriksa jawaban AI sepanjang hari juga menyadarinya?

Malam di Tempe: Operator yang Terlena

Peringatan tahun 1983 itu pernah menjadi kenyataan di sebuah jalan di Arizona.

MINGGU, 18 MARET 2018, PUKUL 21.58. Di Tempe, Arizona, sebuah mobil uji milik unit kendaraan swakemudi Uber, Uber Advanced Technologies Group, menabrak seorang pejalan kaki. Mobil itu sedang dikendalikan sistem mengemudi otomatis yang masih dikembangkan. Di balik kemudi duduk seorang operator, yang tugasnya satu: memantau jalan dan sistem, lalu mengambil alih bila sistemnya keliru.

Pada 19 November 2019, Dewan Keselamatan Transportasi Nasional Amerika Serikat, NTSB, mengadopsi laporan penyelidikannya.[107] Penyebab utamanya, menurut NTSB, adalah kegagalan operator memantau lingkungan berkendara dan kerja sistem otomatis karena ia terdistraksi ponsel pribadinya sepanjang perjalanan.

Kalau laporan itu berhenti di situ, kisahnya sederhana: ada orang yang lalai. Namun laporan itu juga menyebut tiga hal di pihak perusahaan yang ikut menyebabkan tabrakan: prosedur penilaian risiko keselamatan yang tidak memadai, pengawasan yang tidak efektif terhadap para operator, dan tidak adanya mekanisme yang memadai untuk menangani *automation complacency* para operator, yaitu rasa terlena karena sistem otomatis biasanya bekerja dengan baik.[107] Laporan yang sama mencatat bahwa keputusan meniadakan operator kedua menambah beban tugas operator yang tinggal sendirian, dan menyebut perhatian yang lama teralih itu sebagai efek khas dari rasa terlena tersebut.

Bagi saya, kalimat terpenting dalam laporan itu justru tidak menyangkut ponsel. Kalimat terpentingnya adalah pengakuan bahwa rasa terlena bisa diramalkan, dan karena bisa diramalkan, perusahaan wajib

merancang pengaman untuk menghadapinya. Pemeriksa yang berhenti memeriksa bukan kejadian langka yang menimpa orang yang kebetulan lalai. Itu hasil yang wajar dari sistem yang hampir selalu benar, bila pemeriksaannya tidak dirancang untuk mencegah kelengahan tersebut.

Kantor Anda tidak mengoperasikan mobil swakemudi. Namun pola yang sama bisa terjadi pada orang yang setiap hari menyetujui ringkasan kontrak buatan AI, analisis kredit buatan AI, atau jawaban untuk pelanggan yang dibuat AI. Ketika sesuatu lolos, perusahaan mudah sekali mencari kambing hitam, yaitu menyalahkan orang terakhir yang membubuhkan paraf. Pertanyaan yang lebih jujur adalah pertanyaan para penyelidik di Tempe: apa yang sudah dilakukan perusahaan supaya paraf itu tetap berarti?

Ada satu rincian lagi yang patut direnungkan direksi. Mobil uji itu pernah diawasi dua orang, lalu satu. Keputusan seperti itu hampir selalu masuk akal di atas kertas: sistemnya makin baik, biayanya makin ringan, dan orang kedua jarang sekali harus bertindak. Di kantor, keputusan serupa terjadi tanpa disadari. Dulu ada pembuat dan pemeriksa; sekarang AI membuat, satu orang memeriksa, dan orang itu juga diminta menyelesaikan pekerjaannya sendiri. Bebannya bertambah tepat ketika godaan untuk percaya juga bertambah. Mengurangi pemeriksa tidak selalu salah. Yang salah adalah menguranginya tanpa bertanya apa yang akan menjaga perhatian orang yang tersisa.

Cap “Dari AI” yang Membuat Kesalahan Lolos

Ada yang lebih halus daripada rasa bosan: cara kita memandang mesin itu sendiri.

Pada Juni 2026, jurnal *PNAS Nexus* menerbitkan eksperimen terhadap 1.339 guru di Yunani.[108] Rancangannya didaftarkan lebih dulu sebelum

dijalankan. Setiap guru diminta memeriksa sebuah jawaban siswa yang sudah diberi nilai, dan nilai itu sengaja dibuat keliru: ada yang terlalu rendah, ada yang terlalu tinggi. Bedanya hanya pada label. Sebagian guru diberi tahu nilai itu berasal dari rekan guru. Sebagian lagi diberi tahu nilai itu berasal dari sistem AI.

Untuk nilai yang terlalu tinggi, label tidak membawa perbedaan. Untuk nilai yang terlalu rendah, ceritanya lain. Ketika nilai yang terlalu rendah itu berlabel AI, selisih antara nilai akhir guru dan nilai yang adil menjadi 22 persen lebih besar dibanding ketika nilai yang sama berlabel rekan guru. Kesalahan yang sama lebih mudah lolos hanya karena dicap "dari AI".

Yang lebih mengejutkan adalah siapa yang paling sering membiarkan nilai keliru berlabel AI itu lolos. Menurut para peneliti, kecenderungan itu paling kuat pada guru yang lebih muda, berpendidikan tinggi, dan percaya diri dengan teknologi. Padahal ketika ditanya pendapatnya tentang pihak yang memberi nilai, para guru memberi peringkat yang jauh lebih buruk bila pihak itu disebut algoritma. Mereka menilai pihak itu kurang baik, tetapi tetap membiarkan kesalahannya lewat.

Polanya sama dengan temuan di ruang radiologi:[109] saran keliru dari mesin cenderung lolos, bahkan di tangan para ahli, dan bahkan di tangan orang yang menilai sumbernya kurang baik.

Bagi perusahaan, ada satu pelajaran praktis yang murah. Cara sebuah saran disajikan ikut menentukan apakah ia diperiksa. Formulir yang mencetak "Rekomendasi sistem: SETUJU" dengan huruf tebal di bagian atas sudah mendorong pemeriksa untuk setuju sebelum ia membaca baris pertama. Bandingkan dengan formulir yang meminta pemeriksa menuliskan kesimpulannya sendiri dulu, baru menampilkan saran sistem di halaman berikutnya. Isinya sama, alatnya sama, tetapi urutannya

membuat pemeriksa berpikir sendiri lebih dulu sebelum melihat jawaban sistem.

Penjelasan yang Terlalu Meyakinkan

Kalau masalahnya manusia terlalu percaya pada mesin, bukankah jalan keluarnya sederhana? Suruh AI menjelaskan alasannya, lalu latih karyawan agar lebih kritis.

Sayangnya, bukti penelitian menunjukkan bahwa persoalannya tidak sesederhana itu.

Pada November 2024, *Radiology* kembali menerbitkan eksperimen.[110,111] Kali ini pesertanya 220 dokter: 132 radiolog, sisanya dokter penyakit dalam dan dokter gawat darurat. Mereka membaca foto rontgen dada dengan bantuan asisten AI simulasi yang kinerjanya setara para ahli.

Asisten itu memberi saran diagnosis, kadang benar, kadang keliru, selalu dengan penjelasan dalam salah satu dari dua bentuk. Penjelasan "lokal" menandai bagian gambar yang dianggap paling penting. Penjelasan "global" memperlihatkan gambar serupa dari kasus-kasus lain.

Ketika saran AI benar, dokter yang diberi penjelasan lokal menjawab benar 92,8 persen, dibanding 85,3 persen pada dokter yang diberi penjelasan global. Angka ini adalah ketepatan diagnosis para dokter, bukan ketepatan AI-nya. Ketika saran AI keliru, ketepatan para dokter jatuh ke 23,6 persen dengan penjelasan lokal dan 26,1 persen dengan penjelasan global. Kira-kira hanya satu dari empat diagnosis yang tetap benar ketika sarannya salah.

Penjelasan membantu ketika AI benar, tetapi tidak menyelamatkan dokter ketika AI keliru. Salah seorang peneliti seniornya, Paul Yi,

mencatat bahwa dokter yang diberi penjelasan lokal cenderung lebih cepat memercayai diagnosis AI, baik ketika saran itu benar maupun ketika salah. Penjelasan yang rapi ternyata bisa membuat kita lebih cepat setuju, bukan lebih teliti.

Bagaimana dengan pelatihan? Sebuah uji acak terhadap 44 dokter, yang semuanya sudah mengikuti pelatihan literasi AI selama 20 jam, menemukan bahwa mereka yang menerima rekomendasi AI yang keliru mencapai ketepatan diagnosis 14 poin persen lebih rendah daripada yang menerima rekomendasi benar.[112] Dua puluh jam pelatihan tidak membuat para dokter kebal.

Ini tidak berarti pelatihan dan penjelasan tidak berguna. Pelatihan tetap perlu, dan penjelasan yang baik tetap membantu ketika AI benar. Yang keliru adalah menganggap keduanya sudah menyelesaikan masalah, lalu berhenti di situ.

Kembali ke satpam tadi: penjelasan dan pelatihan membantu pemeriksa melihat lebih jelas, tetapi tidak membuat orang yang sudah terbiasa melambaikan tangan kembali membuka tas.

Ketika Pemeriksaan Hanya Menjadi Stempel

Ada satu temuan lagi yang terasa berlawanan dengan akal sehat.

Banyak perusahaan meminta pemeriksa bukan hanya menilai, tetapi juga langsung memperbaiki hasil AI. Cara ini tampak efisien: sekali jalan, dua pekerjaan beres.

Sebuah eksperimen acak dengan 2.784 peserta menemukan hal sebaliknya.[113] Ketika pemeriksa diwajibkan memperbaiki sendiri setiap kesalahan AI yang sudah ditandai, mereka justru makin kurang terlibat dalam pemeriksaan dan lebih condong menerima saran yang salah.

Dugaan saya sederhana: memperbaiki itu melelahkan. Menyatakan “ada yang salah” berarti menambah pekerjaan sendiri, dan jauh lebih ringan menganggap semuanya sudah benar.

Orang yang terjebak dalam situasi ini punya julukan yang pas: manajer tukang stempel.[4] Prosesnya masih mewajibkan seseorang menandatangani, tetapi orang itu tidak lagi benar-benar menilai. Ia sekadar membubuhkan stempel pada hasil kerja mesin. Peran seperti ini makin rentan, karena stempel adalah bagian pekerjaan yang paling mudah diotomasi. Persetujuan tanpa penilaian juga tidak melatih siapa pun; orang yang lima tahun hanya menyetujui tidak akan lebih pandai menilai di tahun keenam.

Dari situ muncul pandangan yang keras: untuk alur kerja yang sudah matang dan batasnya jelas, tahap “biar dicek orang dulu” memakan waktu dan gaji, jadi perusahaan perlu bisa menunjukkan kesalahan apa yang dicegahnya; tahap itu tidak otomatis membuat aman. Saya setuju, dengan satu batas yang tegas. Pandangan itu berlaku untuk pemeriksaan rutin pada tugas berisiko rendah. Untuk tindakan kritis, yang tidak bisa dibatalkan dan dampaknya besar, pemeriksaan dua orang tetap wajib, seperti yang akan kita bahas di Bagian Pembahasan "Kala AI Mulai Bertindak". Masalahnya bukan banyak atau sedikitnya pemeriksa, melainkan apakah pemeriksaan itu ditempatkan di tempat yang tepat dan benar-benar terjadi.

Prinsipnya bisa dirangkum dalam satu kalimat: pemeriksaan manusia bukan pengaman otomatis. Pemeriksa yang tidak benar-benar memeriksa memberi rasa aman palsu, sekaligus memperlambat proses.

Lalu bagaimana menjaga agar pemeriksaan tetap terjadi? Saran saya dimulai dari tiga hal.

Pertama, batasi persetujuan manusia hanya untuk tindakan yang benar-benar berdampak tinggi. Bila setiap hasil AI wajib disetujui manusia, jumlah persetujuannya terlalu banyak, dan perhatian pemeriksa habis sebelum sampai ke keputusan yang penting.

Kedua, untuk keputusan yang berat, minta manusia menilai sendiri lebih dulu, sebelum melihat jawaban AI.

Ketiga, ukur proses pemberian persetujuan itu sendiri. Bila seorang pemeriksa menyetujui hampir semua yang lewat, dengan waktu baca yang sangat singkat, itu tanda bahaya. Untuk itu saya merancang tiga kontrol: batasi jumlah persetujuan per orang per jam, periksa ulang sebagian sampel secara independen, dan rotasi pemeriksanya. Yang dipantau ada tiga angka: tingkat persetujuan, lama waktu membaca, dan seberapa sering pemeriksa mengubah hasil AI.

Batas persetujuan per jam mengakui bahwa perhatian manusia punya batas, seperti batas setengah jam tadi, dan bahwa persetujuan yang kesekian ratus dalam satu sore tidak bernilai sama dengan persetujuan pertama. Rotasi memberi mata yang segar pada pekerjaan yang lama-lama terasa sama, tanpa perlu meragukan kejujuran siapa pun. Sampel yang diperiksa ulang secara independen adalah satu-satunya cara mengetahui berapa yang lolos tanpa menunggu pelanggan yang memberi tahu. Ketiganya dipasang karena perhatian manusia pasti menurun, bukan karena pemeriksanya dicurigai.

Angka seperti ini bisa dicatat, dan sebagian perusahaan sudah mencatatnya. Di Bagian Pembahasan "Sudah Investasi AI, Kok Belum Cuan?" kita melihat, menurut data pemakaian tahap awal, para pengembang Telkom sudah menerima 20 sampai 30 persen saran kode pemrograman dari asisten AI.[19] Pelajarannya: berapa persen saran AI yang diterima pemeriksa adalah angka yang bisa dicatat. Perusahaan

yang tidak mencatat angka itu tidak tahu apakah para pemeriksanya masih benar-benar memeriksa.

Lima Minggu Tanpa Temuan

Kembali ke PT Contoh Distribusi, pada pekan setelah percakapan di meja Pak Hendra.

Pemeriksanya seorang staf penagihan senior, orang yang sama yang dulu ikut menulis jawaban benar untuk kumpulan uji. Ia tidak lalai. Pada minggu-minggu pertama, ia menandai beberapa jawaban AI yang keliru dan memperbaikinya. Lalu temuannya menipis. Waktu yang ia habiskan untuk memeriksa 50 jawaban turun dari sekitar dua jam menjadi 20 menit per minggu. Dalam hitungan kasar, dari sekitar dua setengah menit per jawaban menjadi 24 detik. Lima minggu terakhir, ia tidak menemukan satu kesalahan pun.

Bu Rini tidak memanggilnya. Ia memanggil kepala bagian penagihan dan mengusulkan sebuah percobaan kecil. Di antara 50 jawaban AI untuk minggu berikutnya, mereka menyelipkan lima jawaban keliru yang kesalahannya sudah mereka ketahui: tanggal jatuh tempo yang meleset seminggu, kelompok pelanggan yang tertukar, satu jawaban yang menyebut tanggal libur sebagai hari kerja. Pemeriksanya tidak diberi tahu.

Hasilnya diterima pada hari Jumat. Dari lima kesalahan, satu yang tertangkap.

Pak Dedi, kepala penjualan, yang menghadiri rapat sore itu, nyaris marah. “Jadi selama ini pemeriksaannya cuma formalitas?”

“Bukan orangnya yang salah, Pak,” kata Bu Rini. “Rancangannya. Kita memintanya mencari jarum di tumpukan jerami yang hampir selalu kosong. Siapa pun akan berhenti mencari.”

Rancangan itulah yang mereka ubah, dan perubahannya hampir tidak memakan biaya. Jawaban AI kini wajib menyebut pasal kebijakan yang menjadi dasar tanggal jatuh temponya, sehingga pemeriksa cukup mencocokkan, tidak perlu menghitung ulang. Untuk sepuluh dari 50 jawaban, pemeriksa menuliskan dulu tanggal yang menurutnya benar sebelum melihat jawaban AI. Dua staf senior melakukan pemeriksaan secara bergantian. Setiap kesalahan yang ditemukan disebut di rapat bulanan sebagai hasil kerja pemeriksa, bukan dianggap menambah pekerjaannya.

Sebulan kemudian, percobaan diulang dengan lima kesalahan baru. Kali ini empat tertangkap.

Empat dari lima belum sempurna. Tetapi sekarang PT Contoh tahu, dengan angka, berapa banyak kesalahan yang tertangkap pemeriksanya; sebelumnya hal itu hanya diasumsikan. Pak Hendra menambahkan satu baris ke lembar aturan pakai: kesalahan yang diselipkan diuji ulang setiap kuartal.

Ia juga mengubah cara membaca rekap mingguan. Dulu, angka nol di kolom temuan dibaca sebagai kabar baik. Sekarang, angka nol untuk jawaban yang menurut kumpulan uji benar 97 dari 100 dibaca sebagai pertanyaan: apakah AI-nya mendadak membaik, atau pemeriksaannya yang melemah? Bila AI memang membaik, kumpulan uji akan menunjukkannya. Bila tidak, pemeriksaan yang perlu dibenahi. Rekap yang sama, dibaca dengan cara berbeda, kini memberi tahu apakah pemeriksaan masih berjalan.

Gagasan menyelipkan kesalahan yang sudah diketahui juga tidak lahir di PT Contoh. Di pemeriksaan bagasi bandara, mesin sinar-X lazim dilengkapi fungsi *threat image projection*: perangkat lunak yang menyisipkan gambar barang berbahaya fiktif ke citra sinar-X bagasi penumpang, sebelum citra itu tampil di layar petugas. Menurut para peneliti keamanan, fungsi ini menaikkan perhatian dan motivasi petugas, dan sering dipakai untuk mengukur kinerja deteksi mereka.[114] Satpam di pintu bandara, berbeda dengan satpam di pintu pusat perbelanjaan tadi, sesekali diberi tas yang memang berisi sesuatu. Karena itu ia tetap membuka tas.

Empat Jenis Keputusan

Kisah bagian penagihan PT Contoh memperlihatkan cara memperbaiki satu pemeriksaan. Pertanyaan yang lebih besar adalah di mana saja pemeriksaan manusia perlu ada.

Langkah sederhana yang saya sarankan: jangan perlakukan semua keputusan dengan cara yang sama. Pilah dulu keputusannya ke dalam empat jenis.

Jenis pertama, **hanya manusia**. Ini untuk keputusan yang tidak bisa dibatalkan dan membawa tanggung jawab hukum atau pribadi, misalnya keputusan disiplin atau pemutusan hubungan kerja seorang karyawan. AI hanya boleh membantu urusan administrasi. Keputusan dan alasannya dicatat oleh manusia.

Jenis kedua, **manusia dulu**. Ini untuk keputusan berisiko tinggi, keputusan yang berada di luar batas kemampuan AI yang sudah teruji di pembahasan sebelumnya, atau tugas yang penting untuk proses belajar karyawan, misalnya keputusan kredit besar, klaim, dan rekrutmen. Manusia menilai sendiri lebih dulu. AI baru dipakai sesudahnya, sebagai

penantang atau pendapat kedua, dan ketidaksepakatan antara keduanya dicatat.

Jenis ketiga, **AI membantu**. Ini untuk keputusan yang bisa dibatalkan dan berada di dalam atau dekat batas kemampuan AI, misalnya penyusunan draf kontrak standar. AI menyusun pilihan dan draf, manusia memutuskan dan bertanggung jawab atas hasilnya.

Jenis keempat, **AI dulu**. Ini untuk pekerjaan rutin, berisiko rendah, bisa dibatalkan, dan sudah teruji, misalnya memilah *email* atau tiket internal. AI langsung mengeksekusi. Manusia mengaudit sampel secara berkala dan menangani pengecualian.

Bila ragu, pilih jenis yang lebih protektif.

Jenis	Kapan dipakai	Peran manusia dan AI
Hanya manusia	Tidak bisa dibatalkan; membawa tanggung jawab hukum atau pribadi (disiplin, PHK)	Manusia memutuskan dan mencatat alasan; AI untuk administrasi
Manusia dulu	Risiko tinggi, di luar batas AI yang teruji, atau penting untuk belajar (kredit besar, klaim, rekrutmen)	Manusia menilai dulu; AI penantang; ketidaksepakatan dicatat
AI membantu	Bisa dibatalkan, di dalam atau dekat batas AI (draf kontrak standar)	AI menyusun; manusia memutuskan dan bertanggung jawab
AI dulu	Rutin, risiko rendah, bisa dibatalkan, sudah teruji (memilah tiket internal)	AI mengeksekusi; manusia mengaudit sampel dan menangani pengecualian

Pemilahan ini mengubah satu hal penting bagi manajer. Pemeriksaan manusia tidak lagi diwajibkan untuk semua keputusan, sampai semua orang menjadi tukang stempel. Pemeriksaan dipusatkan di tempat yang paling membutuhkannya, lalu dirancang supaya benar-benar terjadi.

Satu catatan untuk pekerjaan inspeksi yang berulang. Dalam sebuah eksperimen dengan 101 peserta pada tugas inspeksi visual yang kritis bagi keselamatan, urutan kerja ketika AI memeriksa pertama, lalu manusia memeriksa ulang, justru memberi keseimbangan terbaik antara hasil pemeriksaan dan keadaan pemeriksanya, seperti kewaspadaan dan pengalaman bekerja sama dengan AI.[115] Ini tidak bertentangan dengan aturan "manusia dulu". Memeriksa barang satu per satu adalah pekerjaan yang berulang, cepat, dan bisa dikoreksi, sehingga lebih dekat ke jenis "AI dulu" dengan audit sampel. Aturan "manusia dulu" berlaku untuk keputusan bernilai tinggi yang jarang terjadi dan berdampak pada orang.

Menilai Dulu, Baru Melihat Saran AI

Mengapa "manusia dulu" berarti manusia benar-benar menilai sebelum melihat jawaban AI? Bukankah cukup menampilkan keduanya berdampingan?

Pada 2021, tiga peneliti menguji pertanyaan ini dengan 199 peserta.[116] Peserta melihat foto sepiring makanan, lalu harus mengganti bahan yang paling tinggi karbohidratnya. Sebuah AI memberi saran. Sebagian peserta menerima saran itu beserta penjelasannya seperti biasa. Sebagian lagi mendapat rancangan yang sengaja memaksa mereka berpikir: saran AI baru muncul bila diminta, atau peserta harus memutuskan sendiri dulu baru melihat saran AI, atau layar menampilkan tulisan "AI sedang memproses gambar" selama 30 detik sebelum saran keluar.

Rancangan yang memaksa peserta berpikir itu secara signifikan mengurangi ketergantungan berlebihan pada AI, dibandingkan dengan pemberian saran yang hanya disertai penjelasan. Namun hasil itu punya

tiga batas. Pertama, ketergantungan itu berkurang, tidak hilang: ketika saran AI keliru, peserta yang bekerja tanpa AI sama sekali tetap lebih baik secara signifikan. Kedua, rancangan yang paling berhasil mengurangi ketergantungan justru mendapat penilaian paling buruk dari pesertanya. Ketiga, manfaatnya paling besar bagi peserta yang memang senang berpikir keras.

Pelajaran yang saya tarik untuk kantor: rancangan layar ikut menentukan apakah pemeriksaan sungguh terjadi. Untuk keputusan jenis “manusia dulu”, saran AI sebaiknya disembunyikan sampai pemeriksa mencatat penilaiannya sendiri.

Siapkan juga diri untuk keluhan. Bila pemeriksa merasa prosesnya lebih lambat dan lebih repot, itu belum tentu tanda kegagalan; keluhan itu mungkin justru tanda mereka mulai benar-benar berpikir.

Sketsa: Analis yang Selalu Setuju

Di Bank Contoh, bank menengah yang kita kenal di Bagian Pembahasan "Langganan AI Dibayar Dolar, tapi Gaji Rupiah" dan Bagian Pembahasan "Batas Acak Kemampuan AI", aturan itu diuji pada berkas kredit. Bank, tokoh, dan angkanya rekaan, dan kesamaan dengan pihak nyata tidak disengaja. Yang nyata adalah latar belakangnya: bank-bank di Indonesia memang sudah memakai AI dalam analisis kredit, misalnya untuk penilaian agunan yang sebelumnya banyak dikerjakan pihak ketiga.[55]

Setahun terakhir, Bank Contoh memakai AI untuk menyusun memo analisis dan rekomendasi bagi permohonan kredit usaha. Analis membaca memo itu, lalu menyetujui atau menolak rekomendasinya. Sekitar 1.200 berkas lewat setiap bulan.

Bu Ratna, direktur keuangan, tidak menemukan masalahnya di laporan kredit macet. Ia menemukannya di dasbor yang jarang dibuka orang. Di sana ada dua angka yang berdampingan. Pertama, analis setuju dengan rekomendasi AI pada 99 persen berkas. Kedua, waktu yang mereka habiskan per berkas terus memendek dari bulan ke bulan.

“Kalau analis kita setuju 99 persen,” tanya Bu Ratna kepada Pak Yusuf, direktur teknologi, “apa bedanya dengan tidak punya analis?”

Pak Yusuf tidak menjawab dengan memecat siapa pun. Ia memilah. Berkas di atas plafon tertentu, sekitar sepersepuluh dari seluruh berkas atau kurang lebih 120 berkas sebulan, dipindahkan ke jenis “manusia dulu”. Untuk berkas-berkas itu, layar analis kini tidak menampilkan rekomendasi AI sampai analis mencatat penilaian dan alasannya sendiri. Baru setelah itu memo AI terbuka, dan bila keduanya berbeda, perbedaannya dicatat. Sisanya, sekitar 1.080 berkas, tetap di jenis “AI membantu”, dengan lima persen sampel, atau sekitar 54 berkas sebulan, diperiksa ulang oleh analis kedua yang tidak tahu siapa penyetuju pertamanya.

Perubahan itu tidak disambut tepuk tangan. Dalam rapat dengan kepala unit kredit, seorang analis senior bertanya terus terang mengapa ia harus menulis penilaiannya sendiri, padahal memo AI sudah lengkap dan hampir selalu benar. Pak Yusuf menjawab dengan pertanyaan balik: bila memo itu hampir selalu benar, bagaimana kita tahu kapan ia tidak benar, kalau tidak ada yang menilai tanpa melihatnya lebih dulu? Untuk berkas kecil, katanya, bank memang memilih percaya pada mesin dan memeriksa sampelnya. Untuk berkas besar, bank membayar analis justru untuk pendapatnya sendiri.

Bulan pertama, dua hal terjadi. Para analis mengeluh prosesnya lebih lambat, persis seperti peserta eksperimen foto makanan tadi. Selain itu,

untuk 14 dari 120 berkas besar, sekitar satu dari sembilan, penilaian analis berbeda dengan rekomendasi AI. Tidak semua perbedaan itu berarti analisnya yang benar; beberapa kali, setelah dibahas, rekomendasi AI ternyata lebih tepat. Tetapi sekarang Bank Contoh punya sesuatu yang tidak pernah dimilikinya: catatan tentang di mana manusia dan mesin tidak sepakat, dan siapa yang lebih sering benar untuk jenis berkas apa.

Bu Ratna memasukkan dua angka baru ke dasbor: tingkat ketidaksepakatan dan hasil pemeriksaan ulang sampel. Ketika angka setuju kembali mendekati 99 persen di sebuah cabang, itu bukan lagi kabar baik otomatis. Itu alasan untuk bertanya. Ketika angka ketidaksepakatan untuk suatu jenis berkas turun mendekati nol selama berbulan-bulan, dan pemeriksaan ulang sampel tidak menemukan masalah, itu juga alasan untuk bertanya, tetapi dengan arah sebaliknya: mungkin jenis berkas itu sudah layak dipindahkan ke “AI membantu”, sehingga waktu analis bisa dipakai untuk berkas yang lebih sulit. Pembagian jenis keputusan ini, sama seperti hasil kumpulan uji di pembahasan sebelumnya, perlu ditinjau berkala: ada jenis berkas yang perlu diperketat, ada pula yang boleh dilonggarkan.

Penjelasan yang Mudah Dicek

Penjelasan AI tidak menyelamatkan dokter ketika AI keliru. Apakah itu berarti penjelasan memang tidak berguna?

Helena Vasconcelos dan rekan-rekannya memberi jawaban yang lebih bernuansa. Dalam lima studi dengan 731 peserta, mereka meminta orang mencari jalan keluar dari labirin dibantu AI simulasi.[117,118] Menurut mereka, orang memilih secara strategis apakah akan memeriksa penjelasan AI atau tidak, seperti menilai untung dan rugi: seberapa repot

memeriksanya, dan seberapa besar manfaatnya. Pada labirin yang sulit, penjelasan sederhana yang langsung menandai jalannya mengurangi ketergantungan berlebihan, sedangkan petunjuk belokan demi belokan yang rumit tidak berpengaruh. Imbalan uang untuk jawaban yang benar juga mengurangi ketergantungan. Saran mereka bagi perancang sistem: buat penjelasan yang membuat kesalahan AI mudah terlihat.

Di kantor, saran itu bisa diterjemahkan menjadi dua langkah.

Pertama, minta AI menunjukkan sumbernya, bukan menjelaskan panjang lebar. Ringkasan kontrak menyebut pasal dan halamannya. Jawaban soal kebijakan menautkan dokumen aslinya. Jawaban jatuh tempo di PT Contoh menyebut pasal kebijakannya. Pemeriksa cukup mencocokkan, tidak perlu membaca ulang semuanya.

Kedua, buat temuan kesalahan bernilai bagi pemeriksa. Pemeriksa yang menemukan kesalahan AI sebaiknya dihargai, bukan dibebani pekerjaan tambahan. Ingat eksperimen 2.784 peserta tadi: ketika pemeriksa wajib memperbaiki setiap kesalahan yang sudah ditandai, mereka justru lebih condong menerima saran yang salah.

Penjelasan yang keliru pun punya bahayanya sendiri; contohnya ada di lantai pabrik, di Bagian Pembahasan "Alarm yang Diabaikan".

Empat Hal di Layar Pemeriksa

Bila semua temuan tadi dikumpulkan, ada satu kesimpulan yang jarang dibicarakan di rapat pengadaan AI: layar tempat pemeriksa bekerja ikut menentukan apakah pemeriksaan terjadi. Layar itu biasanya dirancang vendor atau tim TI, jarang ditanyakan oleh pemilik proses. Saran saya, sebelum sebuah alat AI dipakai untuk pekerjaan yang harus disetujui manusia, pemilik prosesnya memeriksa empat hal di layar itu.

Pertama, bagaimana saran AI diberi label. Eksperimen guru di Yunani menunjukkan bahwa cap “dari AI” saja sudah membuat kesalahan lebih mudah lolos. Saran yang tampil sebagai vonis bercetak tebal di bagian atas layar mendorong orang untuk setuju sebelum membaca.

Kedua, kapan saran AI muncul. Untuk keputusan jenis “manusia dulu”, saran sebaiknya baru muncul setelah pemeriksa mencatat penilaiannya sendiri, seperti di Bank Contoh. Untuk jenis lain, menampilkannya sejak awal boleh saja.

Ketiga, apa yang perlu dicocokkan. Bila layar menampilkan sumber, seperti pasal kebijakan atau halaman kontrak, pemeriksa bisa mencocokkan dalam hitungan detik. Bila layar hanya menampilkan penjelasan yang panjang dan rapi, eksperimen rontgen dada tadi menunjukkan bahwa kerapian itu bisa membuat orang lebih cepat setuju.

Keempat, apa yang tercatat. Layar yang baik mencatat apakah pemeriksa mengubah sesuatu, berapa lama ia membaca, dan apa alasannya bila ia tidak setuju. Tanpa catatan itu, tiga kontrol tadi tidak memiliki data yang diperlukan, dan perusahaan tidak akan pernah tahu apakah pemeriksaannya masih berjalan.

Menanyakan keempatnya sebelum kontrak ibarat sedia payung sebelum hujan: murah bila ditanyakan sebelum kontrak ditandatangani, dan mahal bila baru disadari setelah ribuan persetujuan berlalu.

Beban, Waktu, dan Tanggung Jawab

Eksperimen-eksperimen tadi masing-masing menguji satu rancangan. Bagaimana gambarannya bila banyak studi dikumpulkan sekaligus?

Pada 2012, sebuah tinjauan sistematis di *Journal of the American Medical Informatics Association* menyaring 13.821 makalah tentang bias otomasi, dan 74 di antaranya memenuhi syarat untuk ditinjau.[104] Tinjauan itu mencatat faktor lingkungan yang memperbesar bias otomasi: beban kerja, kerumitan tugas, dan tekanan waktu. Tinjauan yang sama juga mencatat peredamnya, antara lain pelatihan, penegasan tanggung jawab pemakai, dan rancangan tampilan seperti letak saran di layar.

Apakah ini bertentangan dengan temuan bahwa dua puluh jam pelatihan tidak membuat dokter kebal? Tidak. Pelatihan meredam, tidak menghapus. Karena itu, usul saya, temuan tinjauan itu diterjemahkan menjadi tiga langkah yang melengkapi pelatihan.

Pertama, jangan biarkan antrean pemeriksaan tanpa batas. Batas jumlah persetujuan per orang per jam yang tadi kita bahas adalah salah satu jawabannya.

Kedua, beri waktu yang cukup untuk keputusan yang berat. Target kecepatan bagi pemeriksa layak ditinjau ulang, terutama untuk keputusan jenis “manusia dulu”. Dasbor Bank Contoh yang menampilkan waktu per berkas yang makin pendek sebagai prestasi sedang memberi insentif yang keliru.

Ketiga, tegaskan tanggung jawab. Pemeriksa yang menyetujui hasil AI ikut bertanggung jawab atas hasil itu, dan nama serta alasannya tercatat. Catatan itu tidak dimaksudkan untuk menghukum. Gunanya mengingatkan bahwa paraf punya arti.

Benarkah Manusia Selalu Terlalu Percaya?

Sampai di sini, mungkin muncul kesan bahwa manusia selalu terlalu percaya pada AI. Itu juga tidak sepenuhnya benar.

Ingat uji klinis dokter di pembahasan sebelumnya: AI yang bekerja sendiri justru mendapat skor lebih tinggi daripada dokter yang dibantu AI, dan tim penelitinya menduga banyak dokter tidak memasukkan saran model ke pertimbangan mereka. Masalahnya bisa datang dari dua arah. Kadang manusia terlalu percaya pada AI yang keliru. Kadang manusia tidak cukup percaya pada AI yang benar.

Penelitian lain dari ruang radiologi memperlihatkan sisi kedua itu dengan jelas. Empat peneliti menguji 227 radiolog profesional dengan kasus dari 324 rekam medis pasien yang diduga punya kelainan dada.[119] Temuan pertama mereka: memberi radiolog saran dari AI tidak menaikkan kinerja rata-rata, sedangkan memberi informasi konteks tentang pasien justru menaikkan kinerja.

Mengapa saran AI tidak membantu? Menurut para peneliti, radiolog kurang memperhitungkan saran AI, dan memperlakukan penilaian sendiri serta saran AI seolah-olah tidak saling berkaitan. Mereka tidak terlalu percaya. Mereka kurang percaya, pada saat AI sedang benar.

Jadi tujuan kita bukan membuat karyawan curiga pada semua jawaban AI. Tujuannya adalah kepercayaan dengan takaran yang tepat: percaya pada AI untuk tugas yang terbukti mampu dikerjakannya dengan baik, dan waspada pada tugas yang terbukti menjadi kelemahannya. Takaran itu tidak bisa ditentukan dari perasaan atasan. Ia hanya bisa ditentukan oleh data, yaitu peta batas kemampuan AI yang kita susun di pembahasan sebelumnya, dan oleh catatan ketidaksepakatan seperti milik Bank Contoh.

Membagi Kasus antara Manusia dan AI

Temuan kedua dari studi radiolog tadi lebih menarik lagi bagi manajer.

Menurut perhitungan para peneliti, rancangan kerja sama terbaik antara manusia dan AI membagi kasus ke radiolog saja atau ke AI saja, dan jarang sekali ke radiolog yang dibantu AI. Untuk sebagian besar pasien, keputusan terbaik dibuat oleh salah satunya, bukan oleh keduanya bersama-sama.[119]

Jadi, meminta manusia memeriksa ulang hasil AI di setiap kasus belum tentu cara terbaik. Menurut studi ini, hasilnya bisa lebih baik bila setiap kasus diserahkan dengan tegas kepada salah satunya.

Dalam praktik, pembagiannya bisa mengikuti peta batas kemampuan di Bagian Pembahasan "Batas Acak Kemampuan AI" dan empat jenis keputusan tadi. Kasus yang terbukti berada di dalam batas, dengan akibat ringan bila keliru, dikerjakan AI, lalu sampelnya diaudit. Kasus yang sulit atau berat akibatnya dikerjakan manusia sepenuhnya, dengan waktu dan perhatian yang cukup. Itulah yang dilakukan Pak Yusuf ketika memisahkan 120 berkas besar dari 1.080 berkas lainnya.

Studi ini memberi pertanyaan yang layak diajukan setiap direksi: di mana pemeriksaan manusia benar-benar mencegah kerugian, dan di mana ia hanya menambah tanda tangan tanpa penilaian?

Ringkasan rancangan pemeriksaan yang dibahas di bab ini, beserta buktinya:

Rancangan	Bukti	Temuan
Menilai dulu, saran AI bila diminta, atau jeda 30 detik	Eksperimen 199 peserta, tugas foto makanan	Ketergantungan turun, tidak hilang; paling tidak disukai
Penjelasan yang mudah dicek	Lima studi, 731 peserta, tugas labirin	Membantu bila memeriksa lebih murah daripada langsung percaya
Menilai terpisah dari memperbaiki	Eksperimen 2.784 peserta	Koreksi wajib justru menurunkan keterlibatan
Mengurangi beban dan tekanan waktu, menegaskan tanggung jawab	Tinjauan 74 studi (2012)	Beban kerja, kerumitan tugas, dan tekanan waktu memperbesar bias otomasi
Membagi kasus ke manusia saja atau AI saja	227 radiolog	Mayoritas pasien paling baik diputuskan salah satunya
Kesalahan yang sengaja diselipkan	Praktik lazim di pemeriksaan bagasi bandara	Menaikkan perhatian petugas dan dipakai mengukur kinerjanya

Karena sebagian besar bukti di atas datang dari eksperimen, bukan dari kantor yang memakai AI berbulan-bulan, setiap perusahaan sebaiknya menguji rancangan ini di tempatnya sendiri, mulai dari satu unit.

Satu Eksperimen Kecil: Selipkan Kesalahan yang Sudah Diketahui

Bagaimana Anda tahu apakah para pemeriksa di perusahaan Anda benar-benar memeriksa?

Caranya sudah kita lihat di bagian penagihan PT Contoh dan di pemeriksaan bagasi bandara. Saran saya, mulailah dari ukuran yang kecil: satu unit, dua minggu, dua puluh kasus, dengan tiga sampai lima kesalahan yang diselipkan.

Langkahnya bisa dijalankan tanpa anggaran khusus.

Pertama, pilih satu alur kerja yang hasil AI-nya diperiksa manusia: ringkasan kontrak, jawaban untuk pelanggan, atau analisis kredit. Pilih yang tingkat persetujuannya paling tinggi; di sanalah pemeriksaan paling mungkin sudah berubah menjadi stempel.

Kedua, siapkan contoh jawaban keliru yang sudah diketahui jawaban benarnya. Kesalahan itu disusun oleh orang yang paham pekerjaannya, misalnya penulis kumpulan uji di pembahasan sebelumnya, dan harus jenis kesalahan yang memang pernah atau mungkin terjadi, bukan jebakan yang mustahil ditemukan.

Ketiga, selipkan dengan sepengetahuan atasan, tanpa diumumkan kepada pemeriksa, lalu hitung berapa yang tertangkap.

Keempat, ubah rancangannya, bukan orangnya. Bila banyak yang lolos, perbaiki hal-hal yang dibahas di bab ini: sumber yang mudah dicek, penilaian lebih dulu, beban yang dibatasi, rotasi, dan penghargaan bagi temuan. Tujuannya rancangan yang tahan banting: pemeriksaan tetap terjadi walaupun pemeriksanya lelah atau berganti.

Kelima, ulangi. Satu kali uji hanya memberi potret. Uji berkala memberi tahu apakah perbaikannya bertahan. Untuk inspeksi visual di pabrik, khususnya pada bagian produk yang kritis bagi keselamatan, saya menyarankan uji yang sama dengan cacat yang sengaja diselipkan.

Eksperimen seperti ini perlu dijalankan dengan hati-hati. Tujuannya menguji sistem, bukan menjebak atau menghukum orang. Bu Rini tidak memanggil pemeriksanya untuk ditegur; ia memanggil kepala bagian untuk mengubah rancangan. Hasilnya dipakai untuk memperbaiki cara kerja: mengurangi beban persetujuan, memisahkan tugas menilai dari memperbaiki, meminta AI menunjukkan sumbernya, atau memindahkan sebuah keputusan ke jenis yang lebih tepat. Saran saya, beri tahu tim bahwa uji semacam ini akan dilakukan secara berkala, tanpa menyebut kapan dan di berkas mana. Tujuannya menjaga kewaspadaan.

Hasilnya bisa melegakan. Bisa juga mengejutkan, seperti satu dari lima di PT Contoh. Namun setidaknya Anda akan tahu, dengan angka, apakah kalimat “tetap ada manusia yang memeriksa” di perusahaan Anda benar-benar bermakna.

Satu Angka untuk Diingat: Setengah Jam

Bila hanya satu angka dari bab ini yang Anda bawa ke rapat berikutnya, bawalah angka yang paling tua: setengah jam.

Itulah batas perhatian visual yang efektif, menurut makalah tahun 1983 tadi, bahkan bagi orang yang sangat termotivasi, ketika yang diawasinya jarang berubah. Angka itu tidak lahir dari penelitian tentang AI. Tetapi ia menjelaskan hampir semua kisah di bab ini. Radiolog senior yang mengikuti saran keliru. Operator di Tempe yang perhatiannya teralih sepanjang perjalanan. Staf penagihan yang lima minggu tidak menemukan apa-apa. Analis kredit yang setuju 99 persen. Tidak ada di

antara mereka yang jahat atau bodoh. Mereka manusia yang diminta mengawasi sesuatu yang hampir selalu benar, lebih lama daripada kemampuan manusia untuk terus memusatkan perhatian.

Karena itu, setiap kali sebuah proposal AI menjawab risiko dengan kalimat “nanti ada manusia yang memeriksa”, ajukan tiga pertanyaan lanjutan. Memeriksa apa, dan untuk keputusan jenis apa? Berapa lama dalam sehari, dan dengan berapa banyak yang harus disetujui? Bagaimana kita tahu pemeriksaannya masih terjadi bulan depan?

Kalimat di awal bab ini tampak seperti paradoks: semakin canggih sistemnya, semakin penting peran manusianya. Ironinya, semakin canggih sistemnya, semakin mudah pula manusia berhenti memperhatikan. Tugas pemimpin adalah merancang pekerjaan supaya, ketika sistemnya makin canggih, manusia yang mengawasinya tidak berhenti memperhatikan.

Kabar baiknya, rancangan itu jarang mahal. Di PT Contoh, yang berubah hanya cara jawaban disajikan, urutan kerja pemeriksa, jadwal bergantian, dan cara rapat bulanan memperlakukan temuan. Di Bank Contoh, yang berubah hanya kapan memo AI terbuka di layar dan dua angka baru di dasbor. Tidak ada yang membeli alat baru. Yang dibutuhkan hanya kesediaan untuk tidak puas dengan angka nol.

Sebelumnya, kita telah memetakan di mana AI bisa dipercaya. Kini, perhatian beralih pada manusia yang ditugasi memeriksanya. Setelah itu, pembahasan bergerak menuju risiko jangka panjang: apa yang terjadi pada kemampuan manusia itu sendiri bila setiap hari ia bekerja ditemani mesin yang selalu punya jawaban.

Panduan Daftar Pustaka

Halaman-halaman sebelum ini adalah isi bab *Bias Otomasi: Sekadar Tukang Stempel* (Bab 6), dengan satu halaman Identitas Dokumen yang disisipkan setelah halaman pembuka bab. Halaman-halaman setelah ini adalah Daftar Pustaka bab tersebut, memuat 22 sumber. Panduan ini menerangkan cara keduanya saling menunjuk.

Penanda di dalam bab

Sumber sebuah pernyataan ditandai dengan angka kecil di ujung kalimat. Angka itu adalah nomor catatan, bukan bagian dari kalimat dan bukan data.

Contoh bentuknya: “Kalimat yang dikutip dari sebuah sumber.[12]” Angka 12 menunjuk ke entri nomor 12 di Daftar Pustaka.

- Kalimat dengan angka kecil memuat pernyataan atau temuan yang diambil dari sumber lain. Sumbernya adalah entri bernomor sama di Daftar Pustaka.
- Kalimat tanpa angka kecil adalah penjelasan, pertimbangan, atau kesimpulan penulis buku.
- Satu kalimat dapat memiliki lebih dari satu nomor. Setiap nomor menunjuk satu sumber.
- Satu nomor menunjuk satu sumber di seluruh buku. Sumber yang sama memakai nomor yang sama di bab mana pun, sehingga nomor di bab ini tidak harus berurutan dari 1.

Membaca Daftar Pustaka

Setiap entri di Daftar Pustaka dimulai dengan nomor catatan, diikuti nama penulis atau lembaga, judul, tahun terbit, dan alamat tautan bila ada. Untuk mengetahui asal sebuah pernyataan, cocokkan nomor kecil di bab dengan nomor di awal entri.

Daftar Pustaka di berkas ini hanya memuat sumber yang dirujuk bab ini, diurutkan menurut nomor: 4, 19, 55, 101–119. Nomor yang tidak ditemukan di sana berarti sumbernya tidak dicantumkan di dokumen ini.

Menyebut tempat sebuah kutipan

Untuk menyebut asal sebuah pernyataan secara lengkap, ada dua keterangan yang saling melengkapi:

1. **Tempat di buku ini:** judul bab dan nomor halaman cetak.
2. **Sumber aslinya:** entri Daftar Pustaka dengan nomor catatan yang sama.

Nomor halaman cetak terletak di bagian atas setiap halaman isi bab, misalnya “Bias Otomasi: Sekadar Tukang Stempel | 107”. Nomor ini berbeda dari urutan halaman pada berkas PDF, karena berkas ini memuat halaman tambahan di luar isi

bab. Identitas Dokumen, Panduan ini, dan Daftar Pustaka tidak bernomor halaman.

Daftar Pustaka

Bab 6: Bias Otomasi: Sekadar Tukang Stempel. 22 sumber.

4. Lee, Kai-Fu. *AI Native*. Crown Currency, 2026.

19. "Transforming Indonesian Businesses for an AI-Powered Future: Success Stories from BUMA, DANA, and TELKOM." Microsoft News Center Indonesia, 7 Mei 2024. https://news.microsoft.com/id-id/2024/05/07/transforming-indonesian-businesses-for-an-ai-powered-future-success-stories-from-dana-buma-and-telkom/.

55. Saputra, Yulian. "AI jadi mesin efisiensi bank, dorong akuisisi nasabah hingga pertumbuhan CASA." Infobanknews, 3 Juni 2026.

101. Bainbridge, Lisanne. "Ironies of Automation." *Automatica* 19, no. 6 (1983): 775–779.

102. Dratsch, Thomas, dkk. "Automation Bias in Mammography: The Impact of Artificial Intelligence BI-RADS Suggestions on Reader Performance." *Radiology* 307, no. 4 (Mei 2023): e222176.

103. Radiological Society of North America. [Siaran pers tentang bias otomasi dalam mamografi]. 2 Mei 2023. https://rsna.org/media/press/i/2432.

104. Goddard, K., A. Roudsari, dan J. C. Wyatt. "Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators." *Journal of the American Medical Informatics Association* 19, no. 1 (2012): 121–127.

105. "Ironies of Automation." blog.acolyer.org, 8 Januari 2020.

106. IQPlus/ANTARA. "MIND ID Perluas Pemanfaatan AI di Rantai Nilai Pertambangan." BCA Sekuritas, 11 September 2026.

107. National Transportation Safety Board. *Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian, Tempe, Arizona, March 18, 2018*. Highway Accident Report NTSB/HAR-19/03. Diadopsi 19 November 2019.

108. Goulas, Sofoklis, Rigissa Megalokonomou, dan Panagiotis Sotirakopoulos. "Why Do Experts Miss AI's Errors? Evidence from a Randomized Labeling Experiment." *PNAS Nexus* 5, no. 6 (Juni 2026): pgag146.

109. Silicon Canals. [Liputan atas makalah Goulas, Megalokonomou, dan Sotirakopoulos]. 5 September 2026.

110. Prinster, Drew, Amama Mahmood, Suchi Saria, Jean Jeudy, Cheng Ting Lin, Paul H. Yi, dan Chien-Ming Huang. "Care to Explain? AI Explanation Types Differentially Impact Chest Radiograph Diagnostic Performance and Physician Trust in AI." *Radiology* 313, no. 2 (2024).

111. Radiological Society of North America. [Siaran pers tentang penjelasan AI dan diagnosis rontgen dada]. 19 November 2024. https://rsna.org/media/press/i/2552.

112. Qazi, I. A., A. Ali, A. U. Khawaja, M. J. Akhtar, A. Z. Sheikh, dan M. H. Alizai. "Automation Bias in Large Language Model Assisted Diagnostic Reasoning among AI-Trained Physicians." medRxiv, 2025. doi:10.1101/2025.08.23.25334280. Pracetak.

113. Beck, J., S. Eckman, C. Kern, dan F. Kreuter. "Bias in the Loop: How Humans Evaluate AI-Generated Suggestions." arXiv:2509.08514, 10 September 2025. Pracetak.

114. Riz à Porta, Robin, Yanik Sterchi, dan Adrian Schwaninger. “Examining Threat Image Projection Artifacts and Related Issues: A Rating Study.” 52nd IEEE International Carnahan Conference on Security Technology (ICCST 2018), Montréal, Oktober 2018.

115. Berretta, S., A. Tausch, F. Bülow, B. Kuhlenkötter, M. Topp, C. Els, C. Peifer, dan A. Kluge. “Human or AI First? A Holistic Perspective on the Sequential Order of Joint Human-AI Inspection Workflows.” *Applied Ergonomics*, 2025. doi:10.1016/j.apergo.2025.104669.

116. Buçinca, Zana, Maja Barbara Malaya, dan Krzysztof Z. Gajos. “To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making.” *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1, Article 188 (April 2021).

117. Vasconcelos, H., M. Jörke, M. Grunde-McLaughlin, T. Gerstenberg, M. S. Bernstein, dan R. Krishna. “Explanations Can Reduce Overreliance on AI Systems During Decision-Making.” *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1, Article 129 (2023).

118. Miller, Katharine. “AI Overreliance Is a Problem. Are Explanations a Solution?” Stanford HAI, 13 Maret 2023.

119. Agarwal, Nikhil, Alex Moehring, Pranav Rajpurkar, dan Tobias Salz. “Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology.” NBER Working Paper 31422, Juli 2023, direvisi Agustus 2026.