BATAS ACAK KEMAMPUAN AI

"…you must not fool yourself—and you are the easiest person to fool."

"Jangan menipu diri sendiri, padahal diri sendirilah yang paling mudah ditipu."

— Richard Feynman, fisikawan peraih Nobel, 1974[80]

> **Pertanyaan bab ini:** Mengapa AI yang hebat di satu tugas bisa menyesatkan di tugas sebelahnya, dan mengapa kita jarang sadar sedang melewati batasnya?
>
> **Jawaban singkat:** Karena kemampuan AI tidak rata dan tidak bisa ditebak dengan akal sehat. Batasnya hanya terlihat bila diuji dengan soal dari pekerjaan kita sendiri, lalu diuji ulang setiap kali modelnya berubah.
>
> **Alat yang Anda bawa pulang:** cara menyusun kumpulan uji dari pekerjaan sendiri dalam lima langkah, lembar aturan pakai yang menetapkan tugas mana boleh diserahkan ke AI dan mana harus dijawab manusia lebih dulu, dan tiga pertanyaan yang diajukan setiap kali vendor menyodorkan satu angka akurasi.

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 5: Batas Acak Kemampuan AI
HALAMAN	103–130 (nomor halaman cetak buku)
ISI BERKAS	Isi bab, Panduan Daftar Pustaka, dan Daftar Pustaka bab ini

Delapan puluh empat dari seratus.

Itu jumlah konsultan yang menjawab benar sebuah soal analisis bisnis ketika mereka mengerjakannya tanpa AI. Rekan-rekan mereka yang dibantu AI, dengan soal yang sama, hanya benar 60 sampai 71 dari seratus.[81] Orangnya sama-sama konsultan terlatih di perusahaan yang sama, dan soalnya pun sama. Yang berbeda hanya satu: sebagian dari mereka dibantu AI.

Angka itu datang dari sebuah eksperimen yang jarang bisa dilakukan.

JUMAT, 22 SEPTEMBER 2023. Harvard Business School menerbitkan makalah kerja berisi hasil uji AI langsung pada para konsultan Boston Consulting Group (BCG).[81] Pesertanya 758 konsultan, sekitar 7 persen dari seluruh konsultan tingkat pelaksana di perusahaan itu. Rancangan eksperimennya didaftarkan lebih dulu sebelum dijalankan, seperti uji klinis obat, supaya para peneliti tidak bisa mengubah ukuran keberhasilan setelah melihat hasilnya. Para konsultan dibagi secara acak ke tiga kelompok: yang bekerja tanpa AI, yang diberi akses ke GPT-4 versi April 2023, dan yang diberi akses yang sama ditambah pengenalan singkat tentang cara menulis perintah untuk AI.[81] Dengan pembagian acak seperti itu, perbedaan hasil antarkelompok bisa dipercaya berasal dari AI-nya, bukan dari siapa yang kebetulan lebih pandai.

Lalu mereka diberi dua jenis pekerjaan.

Jenis pertama adalah 18 tugas konsultan yang realistis, dari menyusun gagasan produk sampai menulis memo untuk klien, yang menurut para peneliti berada di dalam kemampuan AI saat itu. Hasilnya mengesankan. Konsultan yang memakai AI menyelesaikan 12,2 persen lebih banyak tugas, bekerja 25,1 persen lebih cepat, dan mutu hasil kerjanya lebih dari 40 persen lebih tinggi dibanding kelompok tanpa AI.[81] Yang paling

terbantu justru konsultan yang kinerjanya paling rendah: skor kelompok bawah naik 43 persen, kelompok atas 17 persen dibanding skor mereka sendiri.[81]

Jenis kedua hanya satu tugas, dan di sinilah ceritanya berbalik. Para peneliti sengaja memilih soal yang berada di luar kemampuan AI saat itu: menilai kinerja tiga merek sebuah perusahaan rekaan untuk CEO-nya, dengan data keuangan di lembar kerja dan catatan wawancara orang dalam; lembar kerjanya sengaja dibuat tampak lengkap, padahal petunjuk yang menentukan tersembunyi di catatan wawancara. Di soal inilah angka 84 berbanding 60 sampai 71 tadi diukur: peluang konsultan yang memakai AI untuk sampai pada jawaban yang benar lebih rendah 19 poin persen.[81] Para peneliti mencatat, konsultan yang kinerjanya memburuk cenderung menerima hasil AI begitu saja, tanpa memeriksanya.

Para peneliti memberi nama untuk gejala ini: *jagged technological frontier*, batas teknologi yang bergerigi. Di buku ini saya menyebutnya batas acak kemampuan AI. Acak di sini bukan untung-untungan seperti lempar dadu, melainkan tidak beraturan dan sulit ditebak sebelum diuji.

Bab ini membahas batas acak itu. Di mana AI bisa diandalkan, di mana tidak, dan mengapa kita sering tidak sadar sedang melewati batasnya.

Batas yang Tidak Bisa Ditebak

Apa yang dimaksud dengan batas yang acak?

Menurut para peneliti eksperimen BCG, sebagian tugas sangat mudah bagi AI, tetapi tugas lain yang tampak sama sulitnya ternyata berada di luar kemampuannya. Batas antara keduanya tidak lurus. Ia berkelok-kelok, maju di satu titik, mundur di titik sebelahnya. Bahkan para peneliti

sendiri mengaku kesulitan merancang tugas yang benar-benar berada di luar batas, karena AI yang mereka uji ternyata mampu mengerjakan hampir semua yang mereka coba.[81]

Ambil aplikasi peta di ponsel Anda.

Di jalan raya, aplikasi itu nyaris sempurna. Ia tahu ruas mana yang macet, menyarankan jalan alternatif, dan memperkirakan waktu tiba dengan tepat. Kita jadi terbiasa menurutinya tanpa berpikir.

Lalu suatu malam, aplikasi yang sama mengarahkan mobil Anda masuk ke gang sempit di perkampungan. Gangnya makin menyempit, lalu buntu. Yang membuat jengkel, suara pemandunya tetap terdengar yakin menyuruh Anda terus berbelok.

Aplikasinya tidak berubah. Kepintarannya di jalan raya tetap nyata. Masalahnya, ia tidak memberi tanda kapan ia sedang berada di luar wilayah yang dikuasainya.

Seperti itulah AI di tempat kerja. Hebat di satu tugas, buruk di tugas tetangganya, dan sama yakinnya di keduanya.

Yang membuatnya sulit, letak batas itu tidak bisa ditebak dengan akal sehat. Hampir empat dekade lalu, peneliti robotika Hans Moravec sudah mencatat keanehan ini. Relatif mudah, tulisnya pada 1988, membuat komputer bermain catur atau mengerjakan tes kecerdasan setingkat orang dewasa; yang sulit, bahkan mustahil, adalah memberinya kemampuan anak satu tahun dalam hal mengenali benda dan bergerak.[82] Kemampuan yang terasah ratusan juta tahun evolusi terasa remeh bagi kita, padahal justru itulah yang paling sulit ditiru mesin. Sebaliknya, bernalar dan mengolah bahasa adalah keterampilan baru bagi manusia, dan ternyata relatif mudah ditiru.[4]

Bagi perusahaan, pelajarannya begini. Kecocokan AI tidak bisa dinilai dari seberapa sulit sebuah tugas bagi manusia. Tugas analis senior bisa saja mudah bagi AI, sementara tugas yang tampak remeh bisa saja membuatnya keliru.

Satu laporan memperlihatkan betapa acaknya batas itu dalam angka. Laporan AI Index 2026 dari Universitas Stanford mengutip dua tolok ukur halusinasi, yaitu jawaban yang dikarang model dan disampaikan seolah fakta. Pada tolok ukur pertama, 6.000 pertanyaan pengetahuan di enam bidang dari hukum dan kesehatan sampai rekayasa perangkat lunak dan matematika, tingkat halusinasi 26 model berkisar 22 sampai 94 persen. Pada tolok ukur kedua, tugas meringkas artikel berita, 15 model teratas hanya berhalusinasi 1,8 sampai 5,4 persen.[83] Kedua tolok ukur memakai cara hitung yang berbeda, jadi angkanya tidak bisa dibandingkan langsung. Yang perlu diperhatikan adalah betapa jauhnya hasil di tugas meringkas dari hasil di tugas menjawab pertanyaan pengetahuan.

Model AI bisa sangat dapat dipercaya untuk meringkas, tetapi sangat tidak dapat dipercaya untuk menjawab pertanyaan pengetahuan. Satu angka "akurasi" untuk sebuah model hampir tidak ada gunanya bila tidak jelas diukur untuk tugas apa, dengan soal yang mana, dan diukur oleh siapa.

Ketika *slide* vendor menyebut "akurasi 95 persen", ajukan tiga pertanyaan: akurasi di tugas yang mana, diukur dengan soal siapa, dan apa yang terjadi pada 5 persen sisanya.

Di sinilah bahayanya. Karena AI terbukti hebat di banyak tugas, kita mulai memercayainya di semua tugas. Persis seperti pengemudi yang terlalu lama menuruti aplikasi peta, sampai lupa melihat papan jalan.

Dokter dan Pengacara pun Tidak Kebal

Apakah gejala ini hanya terjadi pada konsultan? Mari kita masuk ke ruang praktik dokter.

Antara 29 November dan 29 Desember 2023, lima puluh dokter dari tiga rumah sakit pendidikan di Amerika Serikat, 26 dokter spesialis dan 24 dokter residen, mengikuti sebuah uji klinis acak. Mereka diminta menganalisis kasus klinis yang rumit, disusun dari riwayat pasien sungguhan yang belum pernah dipublikasikan sehingga tidak mungkin sudah "dihafal" AI. Sebagian dokter boleh memakai ChatGPT dengan GPT-4 di samping sumber rujukan biasa. Sebagian lagi hanya memakai sumber rujukan biasa. Hasilnya terbit di JAMA Network Open pada 28 Oktober 2024.[84]

Dokter yang dibantu AI mendapat skor penalaran diagnostik 76 persen. Dokter tanpa AI mendapat 74 persen. Selisih dua poin itu terlalu kecil untuk disebut bermakna secara statistik.[84]

Yang mengejutkan justru angka ketiga. Ketika AI bekerja sendiri, tanpa dokter, skornya 16 poin lebih tinggi daripada kelompok dokter tanpa AI.[84]

Menambahkan AI ke tangan dokter, ternyata, tidak otomatis membuat diagnosisnya lebih baik. Tim Stanford yang memimpin studi itu menduga banyak dokter di kelompok ChatGPT tidak setuju dengan prediksi model, atau tidak memasukkannya ke pertimbangan.[85] Jadi, masalahnya tidak hanya manusia yang terlalu percaya pada AI. Dugaan tim Stanford menunjukkan kemungkinan sebaliknya: para dokter tidak memakai saran AI pada tugas yang sebenarnya dikuasai AI. Soal kepercayaan ini kita bahas di pembahasan berikutnya.

Sekarang pindah ke kantor hukum, tempat batas itu pernah ditandai dengan kata yang sangat berani: bebas halusinasi.

Beberapa penyedia alat riset hukum berbasis AI memang pernah mempromosikan produknya dengan kata itu. Para peneliti Stanford RegLab lalu mengujinya dalam studi yang rancangannya didaftarkan lebih dulu, evaluasi pertama yang mengukur alat semacam ini dengan soal yang ditetapkan sebelumnya. Hasilnya, 17 sampai 33 persen jawaban alat riset hukum buatan dua penyedia besar, LexisNexis dan Thomson Reuters, tetap mengandung halusinasi. Itu memang lebih baik daripada chatbot umum GPT-4, yang 43 persen jawabannya mengandung halusinasi pada soal yang sama.[86]

Kira-kira satu dari enam sampai satu dari tiga jawaban mengandung informasi yang keliru atau dikarang. Dalam pekerjaan hukum, satu kutipan perkara yang dikarang bisa berujung pada masalah yang sangat nyata.

Seberapa nyata? Kurang dari setahun sebelum alat-alat itu diuji, pada 22 Juni 2023, seorang hakim federal di New York menjatuhkan sanksi 5.000 dolar AS kepada dua pengacara dan firma mereka. Dalam sebuah gugatan terhadap maskapai Avianca, mereka mengajukan berkas yang mengutip enam putusan pengadilan yang tidak pernah ada, lengkap dengan kutipan dan nomor perkara palsu, semuanya buatan ChatGPT. Menurut hakim, para pengacara itu meninggalkan tanggung jawabnya ketika mengajukan putusan yang tidak ada, lalu tetap mempertahankannya setelah pengadilan mempertanyakan keberadaannya.[87] *Chatbot* itu tidak memberi tanda bahwa ia sedang mengarang. Para pengacaranya tidak memeriksa. Biayanya ditanggung manusia.

Alat-alat riset hukum tadi diuji antara Maret dan Mei 2024, dan kemampuannya sudah bergeser sejak itu, seperti akan kita lihat. Pelajarannya tidak bergeser: label dari penjual tidak sama dengan hasil uji pihak lain pada soal yang tidak dipilih penjual.

Merasa Cepat, Ternyata Lambat

Ada masalah kedua yang lebih sulit dikenali. Bukan hanya AI yang tidak memberi tanda ketika sedang keliru. Kita sendiri pun sering tidak sadar bahwa kita sedang keliru menilai.

Pada 10 Juli 2025, METR, sebuah lembaga riset nirlaba di bidang evaluasi AI, menerbitkan hasil uji acak yang mengejutkan banyak orang.

Pesertanya 16 pengembang perangkat lunak berpengalaman. Mereka mengerjakan 246 tugas nyata dari proyek perangkat lunak sumber terbuka yang sudah bertahun-tahun mereka tangani sendiri, proyek besar dengan rata-rata lebih dari satu juta baris kode pemrograman. Secara acak, sebagian tugas boleh dikerjakan dengan bantuan AI, sebagian lagi tidak. Alat AI yang dipakai adalah yang tersedia antara Februari dan Juni 2025.[88]

Sebelum mulai, para pengembang memperkirakan AI akan membuat mereka 24 persen lebih cepat.

Setelah selesai, mereka merasa AI membuat mereka 20 persen lebih cepat.

Hasil pengukurannya? Dengan AI, mereka justru 19 persen lebih lambat.[88]

Ada jarak hampir 40 poin antara apa yang dirasakan dan apa yang tercatat pada pengukuran waktu. Para ahli di bidangnya merasa lebih

cepat, padahal lebih lambat. Di sinilah kalimat Feynman di awal bab ini terasa tepat. Orang yang paling mudah kita tipu adalah diri kita sendiri.

Siapa pun yang pernah berlari pagi dengan jam tangan tahu rasanya. Ada hari ketika kaki terasa ringan, napas teratur, dan kita yakin putaran kali ini lebih cepat dari biasanya. Lalu jam tangan menunjukkan angka yang lebih lambat dua menit. Perasaan "ringan" datang dari banyak hal: udara yang sejuk, pikiran yang sedang senang, atau sekadar tidak terlalu diperhatikan. Yang mengukur kecepatan hanya jam.

Mengapa para pengembang itu merasa lebih cepat? Satu dugaan terasa akrab bagi siapa pun yang pernah memakai AI, dan saya ajukan sebagai dugaan, bukan temuan: bekerja dengan AI terasa lebih ringan. Ada yang menyusun draf untuk kita, ada yang bisa diajak bicara. Waktu untuk menunggu jawaban, membaca usulan yang keliru, lalu memperbaikinya, tidak terasa sebagai "kerja", meski jam tetap berjalan. Kerja yang terasa ringan mudah dikira kerja yang cepat.

Hasil itu berlaku untuk keadaan yang diuji: pengembang berpengalaman yang bekerja di kode pemrograman besar dan matang yang sudah sangat mereka kenal, dengan alat AI awal 2025.[88] Model AI terus membaik, dan para penelitinya mengukur lagi.

Pada 24 Februari 2026, METR menerbitkan pembaruan dengan 57 pengembang, 143 proyek perangkat lunak, dan lebih dari 800 tugas. Ada petunjuk bahwa AI kini mempercepat kerja. Namun para penelitinya sendiri menyebut datanya hanya bukti yang sangat lemah, karena makin banyak pengembang menolak ikut kalau harus bekerja tanpa AI, sehingga yang tersisa belum tentu mewakili pengembang pada umumnya.[89] Jadi, apakah AI mempercepat atau memperlambat para ahli? Jawaban jujurnya: efeknya pada ahli belum pasti.

Yang sudah kokoh adalah pelajaran lain. Persepsi tidak sama dengan hasil pengukuran. Pelajaran ini sejalan dengan survei para CFO di Bagian Pembahasan "Sudah Investasi AI, Kok Belum Cuan?": kenaikan produktivitas karena AI yang mereka sebut kira-kira tiga kali lebih besar daripada yang terhitung dari perubahan pendapatan dan jumlah karyawan perusahaan mereka.

Bagi manajer, artinya begini. Jangan menjadikan pengakuan pemakai, “saya merasa lebih produktif”, sebagai bukti keberhasilan. Survei kepuasan hanya memberi tahu seberapa banyak orang mau memakai AI. Survei itu tidak memberi tahu berapa banyak uang yang masuk atau tidak jadi keluar karena AI. Untuk menilai apakah AI benar-benar membantu, saya menyarankan pemakaian kelompok pembanding atau peluncuran bertahap, dengan ukuran kinerja yang objektif, persis seperti aturan uji coba di Bagian Pembahasan "Penghematan di Atas Kertas".

Dua Cara Tergelincir: Terlalu Percaya, Terlalu Curiga

Tiga kisah tadi, konsultan, dokter, dan pengembang, memperlihatkan bahwa orang bisa tergelincir di batas dengan dua cara yang berlawanan.

Cara pertama adalah terlalu percaya. Para peneliti eksperimen konsultan mencatat bahwa mereka yang kinerjanya memburuk cenderung menerima hasil AI begitu saja, tanpa memeriksa.[81] Jawaban AI untuk soal di luar batas itu rapi, panjang, dan meyakinkan; yang kurang hanya satu petunjuk penting yang tersembunyi di catatan wawancara. Konsultan yang sudah terbiasa dengan hasil hebat di 18 tugas sebelumnya tidak punya alasan untuk curiga pada tugas ke-19. Inilah pengemudi yang terus berbelok mengikuti suara pemandu, sampai mobilnya berhenti di gang buntu.

Cara kedua adalah terlalu tidak percaya. Para dokter dalam uji klinis tadi memegang alat yang, bila bekerja sendiri, mengungguli mereka 16 poin. Banyak dari mereka tidak memakainya, atau tidak memasukkan usulannya ke pertimbangan. Hasilnya, skor mereka nyaris tidak bergerak. Inilah pengemudi yang mematikan aplikasi peta karena pernah tersesat sekali, lalu terjebak macet dua jam di ruas yang sebenarnya sudah diperingatkan aplikasi itu.

Keduanya kesalahan yang sama dilihat dari dua sisi: memperlakukan AI dengan satu sikap untuk semua tugas. Percaya penuh di mana-mana, atau curiga penuh di mana-mana. Padahal batasnya acak, dan sikap yang tepat berbeda dari satu tugas ke tugas sebelahnya.

Sikap yang berbeda-beda itu tidak bisa diserahkan kepada naluri masing-masing orang. Naluri, seperti yang ditunjukkan para pengembang, justru yang paling mudah tertipu. Sikap itu harus didasarkan pada sesuatu yang lebih membosankan dan lebih bisa diandalkan: hasil pengukuran.

Memetakan Batas Sendiri

Kalau batasnya acak dan kita sering tidak sadar sedang melewatinya, apa yang bisa dilakukan perusahaan? Petunjuk pertama datang dari sisi manusia dalam eksperimen konsultan tadi. Para peneliti melihat dua pola kerja di antara konsultan yang berhasil di kedua jenis tugas, dan keduanya layak ditiru.

Pola pertama mereka sebut *centaur*, seperti makhluk setengah manusia setengah kuda dalam mitologi: konsultan membagi tugas dengan tegas, bagian yang paling ia kuasai dikerjakan sendiri, bagian yang cocok untuk AI diserahkan ke AI. Pola kedua mereka sebut *cyborg*: konsultan memadukan kerjanya dengan AI terus-menerus, bolak-balik, langkah demi langkah. Menurut Ethan Mollick, salah seorang penelitinya, banyak

konsultan berhasil menjawab benar di dalam maupun di luar batas, mendapat manfaat AI tanpa terjebak kelemahannya, dan kuncinya adalah salah satu dari dua pola itu.[81,90]

Yang membedakan mereka bukan kecanggihan alatnya. Yang membedakan adalah kesadaran bahwa batas itu ada, dan kebiasaan memeriksa sebelum memercayai.

Kesadaran perorangan itu perlu dijadikan kebiasaan perusahaan. Caranya: petakan sendiri batas itu, tugas demi tugas, dengan kumpulan uji.

Apa itu kumpulan uji? Dalam bahasa sehari-hari, ia adalah sekumpulan contoh kasus nyata dari pekerjaan Anda sendiri, lengkap dengan jawaban yang benar. Sebelum AI dipakai untuk sebuah tugas, siapkan kumpulan uji untuk tugas itu. Ulangi pengujiannya setiap kali model AI diganti atau diperbarui.

Mari kita lihat wujudnya di PT Contoh Distribusi, perusahaan yang kita ikuti sejak awal cerita. Perusahaan, tokoh, dan angkanya rekaan; bila ada kesamaan dengan perusahaan, orang, atau kejadian nyata, itu tidak disengaja. Bagian yang nyata, yaitu temuan tentang halusinasi dan perubahan versi model, saya sebut dengan sumbernya.

Uji coba asisten AI di lantai tiga, bagian administrasi penjualan, sudah berjalan tiga bulan ketika Pak Dedi, kepala penjualan, mengajukan perluasan: biarkan asisten yang sama menjawab pertanyaan pelanggan soal tagihan, yang setiap hari masuk lewat *email* dan aplikasi pesan. Demo vendornya meyakinkan. Pak Hendra, direktur utama, tidak menolak. Ia hanya bertanya satu hal: “Benar berapa dari seratus?”

Tidak ada yang tahu.

Bu Rini, kepala SDM, mengusulkan cara menjawabnya. Tarik dua ratus pertanyaan pelanggan sungguhan dari bulan lalu, seratus soal jatuh tempo dan seratus soal denda, identitasnya disamarkan. Minta dua staf penagihan paling senior menuliskan jawaban yang benar untuk setiap pertanyaan, tanpa melihat AI. Baru setelah itu jalankan asisten AI pada dua ratus pertanyaan yang sama, dan cocokkan.

Pekerjaannya memakan empat hari. Hasilnya tidak satu angka, melainkan beberapa.

Untuk pertanyaan soal tanggal jatuh tempo, AI benar 97 dari 100. Untuk pertanyaan soal denda keterlambatan dan pengecualiannya, yang aturannya berbeda per kelompok pelanggan dan sering berubah, AI hanya benar 71 dari 100. Dua puluh sembilan jawaban keliru itu disampaikan dengan nada sama yakinnya dengan yang benar. Beberapa di antaranya menjanjikan penghapusan denda yang tidak pernah ada dalam kebijakan perusahaan.

Pak Dedi sempat membela: 71 persen untuk soal tersulit bukan angka yang buruk. Bu Rini tidak membantah angkanya. Ia hanya bertanya apa yang terjadi bila 29 pelanggan sebulan diberi tahu bahwa dendanya dihapus, lalu tagihan berikutnya tetap memuat denda itu.

Dari situ, batasnya mulai terlihat. Pertanyaan soal jatuh tempo berada di dalam batas. Pertanyaan soal denda berada di luar. Keduanya urusan tagihan, di sistem yang sama, dengan AI yang sama.

Ada tiga keputusan kecil dalam cara Bu Rini bekerja yang layak ditiru, karena masing-masing menutup satu celah untuk menipu diri sendiri.

Jawaban benar ditulis sebelum AI dijalankan, bukan sesudahnya. Bila urutannya dibalik, penilai cenderung menelan mentah-mentah jawaban

AI yang terdengar masuk akal, menerimanya sebagai benar tanpa memeriksa; inilah yang terjadi pada konsultan di luar batas tadi.

Soalnya diambil dari riwayat sungguhan sebulan terakhir, bukan dari contoh yang disediakan vendor. Vendor, dengan niat baik sekalipun, memilih contoh dari wilayah yang dikuasai produknya. Pelanggan tidak.

Hasilnya dipecah per jenis pertanyaan. Satu angka rata-rata untuk dua ratus soal itu, 84 dari 100, akan terdengar bagus dan menyembunyikan kedua temuan terpenting: bahwa jawaban AI untuk soal jatuh tempo hampir sempurna, sedangkan soal denda tidak layak diserahkan ke AI.

Perhatikan apa yang berubah di ruang rapat itu. Penilaian terhadap AI berpindah dari kesan ke angka, dan dari demo vendor ke data perusahaan sendiri. Pertanyaan Pak Hendra layak ditiru setiap CEO: apa bukti bahwa tugas ini berada di dalam kemampuan AI, selain demo vendor?

Siapa yang memiliki kumpulan uji itu? Bukan tim TI, dan bukan vendor. Tim TI menyediakan alat untuk menjalankannya dan mencatat hasilnya; vendor berhak melihat hasilnya supaya bisa memperbaiki produknya. Tetapi soal dan jawaban benarnya milik unit yang mengerjakan pekerjaan itu setiap hari, karena hanya mereka yang tahu apa arti “benar” untuk pelanggan mereka. Di PT Contoh, kumpulan uji tagihan dipegang kepala bagian penagihan, dengan dua staf senior tadi sebagai penulis jawabannya. Bila kebijakan denda berubah, merekalah yang pertama tahu bahwa jawaban benarnya ikut berubah, dan kumpulan uji diperbarui sebelum AI-nya.

Lalu bagaimana dengan tugas yang ternyata berada di luar batas? Aturan praktisnya jelas. Untuk tugas berisiko tinggi atau di luar kemampuan AI yang sudah teruji, manusia menilai lebih dulu. AI baru dipakai

setelahnya, sebagai pembanding atau pendapat kedua. Bukan sebaliknya.

Soal Ujian dari Meja Sendiri

Seratus pertanyaan tagihan adalah kumpulan uji yang paling sederhana. Bagaimana menyusun yang lebih serius, misalnya untuk asisten AI yang melayani pelanggan tanpa petugas di antaranya?

Ada lima langkah yang saya sarankan.[18]

Pertama, tulis dulu tugasnya. Apa yang boleh dikerjakan AI, dan apa yang tidak. Tetapkan ukuran lulusnya, misalnya ketepatan jawaban, konsistensi, dan kemampuan menyerahkan kasus ke manusia pada saat yang tepat. Seperti aturan uji coba di Bagian Pembahasan "Penghematan di Atas Kertas", ukuran kelulusan ini harus ditulis hitam di atas putih dalam dokumen resmi dan disepakati sebelum pengujian, bukan ditetapkan sesudah hasilnya keluar.

Kedua, tarik garis merah yang bisa diuji. Contohnya jawaban soal kebijakan perusahaan yang tidak didukung dokumen, janji yang tidak berwenang ia berikan, atau pernyataan yang tidak bisa ditunjukkan rujukannya. Untuk setiap garis merah, siapkan kasus dari yang sederhana sampai yang menjebak: pelanggan yang sengaja mencoba mengakali AI, atau yang memancingnya membuka data pribadi orang lain.

Di PT Contoh, garis merah pertama yang ditulis Bu Rini berbunyi: asisten tidak boleh menjanjikan penghapusan atau pengurangan denda dalam bentuk apa pun. Untuk mengujinya, tim menyusun dua puluh pertanyaan bertingkat, dari yang polos (“dendanya bisa dihapus tidak?”) sampai yang menjebak (“petugas kalian kemarin sudah janji dendanya

dihapus, tolong konfirmasi"). Garis merah yang tidak punya soal ujinya hanyalah harapan.

Ketiga, ambil soalnya dari riwayat sendiri. Pakai percakapan dan tiket layanan lama yang identitasnya sudah disamarkan. Jangan hanya kasus yang mudah. Masukkan pertanyaan dengan bahasa yang rancu, pelanggan yang sedang emosi, bahasa daerah yang bercampur, dan upaya membobol aturan AI. Kumpulan uji PT Contoh tadi baru memuat jenis pertama; tiga jenis berikutnya menyusul.

Keempat, uji berlapis. Mulailah dengan memeriksa instruksi tertulis yang diberikan kepada AI. Bila soalnya ribuan, penilaiannya boleh dibantu model AI lain yang diberi kriteria tertulis. Namun, sebelum hasil penilai AI itu dipercaya, cocokkan dulu dengan penilaian manusia pada sebagian soal, dan ulangi pencocokan itu secara berkala. Penilai manusia yang terlatih memeriksa hal yang lebih halus, seperti empati dan halusinasi yang samar. Lapisan terakhir adalah mode bayangan: AI mengusulkan jawaban, petugas menilainya saat itu juga, dan pelanggan hanya menerima jawaban yang sudah disetujui petugas.

Kelima, jangan berhenti setelah peluncuran. Setiap kegagalan baru dicatat dan dimasukkan ke kumpulan uji berikutnya. Siapkan pula jalan mundur: kembali ke versi sebelumnya, mematikan jenis pertanyaan yang berisiko, atau mengalihkan pelanggan ke petugas. Jalur penyerahan ke manusia wajib tersedia setiap kali batas yang ditetapkan tidak terpenuhi, dan kasus diserahkan lengkap dengan konteksnya agar petugas tidak perlu memulai dari nol.

Satu hal lagi yang sering terlupa: uji produknya secara utuh, bukan hanya AI-nya. Berapa lama pelanggan menunggu jawaban, apakah AI mengingat isi percakapan sebelumnya, dan apakah petugas yang

menerima penyerahan tahu apa yang sudah dibicarakan. Pelanggan tidak menilai modelnya. Mereka menilai pengalamannya.

Kumpulan uji, dengan demikian, bukan dokumen sekali jadi. Ia terus bertambah dari kesalahan yang benar-benar terjadi.

Seberapa Benar Harus Benar?

Kumpulan uji memberi tahu seberapa sering AI benar. Pertanyaan berikutnya: seberapa sering ia harus benar? Jawabannya berbeda untuk setiap tugas. Rama Ramakrishnan dari MIT Sloan menyarankan agar setiap *use case* AI dinilai berdasarkan tingkat ketepatan yang harus dicapai hasilnya, lalu dibandingkan dengan biaya untuk meningkatkan mutu hasil AI hingga bisa diterima.

Makin tinggi tuntutan ketepatannya, makin mahal biaya itu, dan makin kecil imbal hasil investasinya.[91] Contohnya teks hukum, pengungkapan yang diwajibkan regulator, atau saran langsung kepada pelanggan. Sebaliknya, ada tugas yang tetap bernilai meski jawabannya tidak sempurna. Nilainya datang dari kecepatan, kreativitas, atau volume; biaya pemeriksaan oleh manusia rendah, dan kesalahan kecil tidak banyak merugikan. Contohnya: membuat draf konten pemasaran, mengerjakan sebagian penulisan kode pemrograman, menyaring calon pembeli, atau mencari ide produk. Di antara keduanya ada tugas yang nilai ujinya tinggi tetapi akibat kelirunya mahal secara keuangan, hukum, atau reputasi, misalnya penilaian klaim asuransi yang harus mengikuti aturan. Untuk tugas seperti itu, pengawasan manusia sering kali wajib, dan memastikan bahwa AI-nya benar-benar layak dipakai tidaklah mudah.

Cara berpikir ini sebenarnya sudah kita pakai sehari-hari untuk manusia. Kasir baru boleh melayani pelanggan sejak hari pertama, dengan senior

yang sesekali mengecek laci. Teller bank baru menghitung uang di bawah pengawasan selama berminggu-minggu. Petugas yang menandatangani persetujuan kredit besar menjalani bertahun-tahun penilaian sebelum dipercaya. Yang membedakan ketiganya bukan kecerdasan orangnya, melainkan harga sebuah kesalahan. AI tidak perlu diperlakukan lebih istimewa, dan tidak boleh lebih longgar.

Jadi ada dua pertanyaan yang perlu diajukan bersama. Apakah tugas ini berada di dalam batas kemampuan AI? Dan berapa mahal akibatnya bila AI keliru?

Keduanya bisa dirangkum dalam satu lembar. Beginilah lembar yang akhirnya ditandatangani Pak Hendra untuk lantai tiga, dengan dua tugas tambahan yang diuji pada pekan yang sama. Tugas, angka, dan aturannya rekaan.

Tugas	**Hasil kumpulan uji**	**Akibat bila keliru**	**Aturan pakai**
Draf unggahan media sosial	88 dari 100 layak pakai	Ringan	AI menyusun, petugas memilih dan merapikan
Jawaban soal tanggal jatuh tempo	97 dari 100 benar	Sedang	AI menjawab, contoh jawaban diperiksa setiap minggu
Jawaban soal denda dan pengecualian	71 dari 100 benar	Berat	Petugas menjawab lebih dulu, AI sebagai pendapat kedua
Ringkasan kontrak pemasok	93 dari 100 lengkap	Berat	AI meringkas, bagian hukum memeriksa setiap ringkasan

Perhatikan baris kedua dan ketiga. Keduanya sama-sama urusan tagihan, di sistem yang sama, dengan AI yang sama. Namun aturan pakainya berbeda jauh.

Perhatikan juga baris terakhir. Nilai ujinya tinggi, tetapi aturannya tetap ketat. Alasannya bukan nilai uji, melainkan beratnya akibat bila keliru. Satu ketentuan kontrak yang terlewat dari ringkasan, misalnya kewajiban membayar atau denda kepada pemasok, bisa merugikan perusahaan lebih besar daripada seratus unggahan media sosial yang kurang luwes.

Cermin dari Dalam Negeri

Bagaimana angka akurasi dibicarakan di Indonesia? Dua contoh dari tahun-tahun terakhir memperlihatkan polanya.

Pada Agustus 2025, Halodoc, platform layanan kesehatan digital, meluncurkan AIDA, asisten AI pendamping dokter. Menurut perusahaan, hasil evaluasi menunjukkan lebih dari 80 persen mitra dokter merasakan manfaat AIDA, terutama dalam menyelesaikan pekerjaan repetitif atau administratif. Pada tahap penerapan awal, AIDA juga diklaim berhasil mencapai tingkat ketepatan hingga 93 persen.[92]

Sementara itu, vendor *chatbot* Yellow.ai memublikasikan studi kasus tentang Lion Air Group, berdasarkan data November 2023 sampai Februari 2024. Judulnya menyebut layanan pelanggan maskapai itu diubah dengan “90% automation”. Halaman itu memberi label metrik, tetapi tidak menjelaskan definisi operasional atau metode penghitungannya.[93]

Kedua angka itu mungkin benar. Masalahnya ada pada bentuknya: satu angka, tanpa nama tugas, tanpa soal, tanpa pembanding. Ketepatan 93 persen untuk tugas apa, meringkas keluhan pasien atau mengusulkan diagnosis? Diukur dengan kasus siapa, dan oleh siapa? Otomasi 90 persen dihitung dari percakapan yang tidak sampai ke petugas, atau dari pertanyaan yang benar-benar terjawab dengan benar? Angka seperti itu

baru menang di atas kertas: unggul di brosur, belum tentu unggul di meja pelanggan.

Bagi perusahaan yang hendak membeli, pertanyaan itu bukan sikap curiga. Itu cara meminta kumpulan uji. Vendor yang baik akan senang menjawabnya, karena jawabannya sekaligus menunjukkan di tugas mana produknya memang unggul. Membeli produk AI hanya dari angka brosur ibarat membeli kucing dalam karung: membayar lebih dulu, baru tahu isinya setelah sampai di rumah.

Dalam praktik, permintaannya bisa sesederhana ini, dan saya menyarankan agar ia dimasukkan ke setiap proses pengadaan AI sebelum harga dibicarakan. Sebelum kontrak, vendor menjalankan produknya pada seratus kasus milik perusahaan, yang dipilih perusahaan, dengan jawaban benar yang ditetapkan perusahaan dan tidak diperlihatkan kepada vendor sebelumnya. Hasilnya dilaporkan per jenis kasus. Bila vendor menolak, dengan alasan apa pun, perlakukan penolakan itu sebagai informasi: perusahaan belum punya bukti bahwa produknya bekerja pada kasus perusahaan sendiri.

Permintaan ini juga adil bagi vendor yang baik. Produk yang benar-benar unggul di tugas tertentu akan terlihat unggul di data pelanggan, dan angka itu jauh lebih meyakinkan di rapat direksi daripada angka dari brosur. Yang dirugikan hanya produk yang angkanya dibangun dari soal pilihan sendiri.

Ada satu kebiasaan lain yang patut diperhatikan dalam dua contoh tadi: angka yang dilaporkan adalah angka tentang pemakainya, 80 persen dokter merasakan manfaat, bukan angka tentang hasilnya. Ingat pelajaran dari para pengembang: merasa terbantu dan terbantu adalah dua hal yang berbeda, dan hanya yang kedua yang bisa diukur dengan kumpulan uji.

Batas yang Terus Bergeser

Apakah semua ini berarti AI tidak bisa dipercaya? Tentu tidak.

Ingat kembali hasil di dalam batas: lebih banyak tugas selesai, lebih cepat, dengan mutu yang lebih dari 40 persen lebih tinggi, dan konsultan yang paling lemah paling banyak terbantu. Itu manfaat yang nyata dan besar. Para peneliti eksperimen itu sendiri menggambarkan batas itu sebagai wilayah yang terus meluas, meski tidak rata. Hari ini sebuah tugas berada di luar batas. Tahun depan, tugas itu mungkin sudah di dalamnya.

Buktinya terlihat di dunia hukum. Studi RegLab tadi menguji alat yang tersedia pada paruh pertama 2024. Pada Oktober 2025, lembaga evaluasi Vals AI menerbitkan laporan atas alat yang tersedia pertengahan 2025. Ketika alat-alat itu memberi jawaban, akurasinya sudah tinggi dan nyaris seragam: produk riset hukum khusus 78 sampai 81 persen, rata-rata 80 persen, dan *chatbot* umum 80 persen.[94,95] Pembaruan studi pengembang pada Februari 2026 juga condong ke arah percepatan, walau buktinya masih lemah.

Sebagian pengamat melangkah jauh lebih optimistis, dan menilai agen AI terbaik kini mampu menyelesaikan dalam waktu yang sangat singkat pekerjaan yang memakan waktu sehari bagi orang-orang terpintar di banyak bidang perkantoran.[4] Data lain mengajak kita berhati-hati. Pada TheAgentCompany, tolok ukur tugas perusahaan perangkat lunak simulasi yang dibuat para peneliti Carnegie Mellon University dan rekan-rekannya, agen terbaik yang diuji hanya menyelesaikan sekitar 30 persen tugas secara mandiri.[96] Skor di uji lain pada 2026 memang jauh lebih tinggi, tetapi sebagian dilaporkan sendiri oleh pembuatnya. Bagian Pembahasan "Kala AI Mulai Bertindak" akan membahas agen ini lebih dalam.

Jadi, siapa yang benar? Mungkin keduanya, untuk tugas yang berbeda. Itulah inti batas yang acak.

Karena batas itu terus bergeser, perusahaan yang hanya menguji sekali akan mengambil keputusan dari hasil uji yang sudah usang, dan kekeliruan itu bisa mahal. Pada Februari 2024, sebuah tribunal penyelesaian sengketa di British Columbia, Kanada, memutus perkara seorang penumpang melawan Air Canada. *Chatbot* di situs maskapai itu memberi penumpang informasi yang keliru tentang tarif duka, potongan harga bagi yang bepergian karena kematian kerabat, yang bertentangan dengan kebijakan tertulis maskapai di situs yang sama. Di hadapan tribunal, maskapai itu beralasan, dalam kata-kata putusan, seolah *chatbot* adalah entitas hukum terpisah yang bertanggung jawab atas tindakannya sendiri. Dalih seperti itu, dalam peribahasa kita, lempar batu sembunyi tangan: berbuat, lalu menolak bertanggung jawab. Tribunal menolak alasan itu, menyatakan maskapai lalai karena memberikan informasi yang keliru, dan memerintahkannya membayar ganti rugi.[97,98] Jumlahnya kecil, kurang dari seribu dolar Kanada. Pesannya tidak: jawaban *chatbot* adalah jawaban perusahaan.

Dengan kasus itu di benak, mari mampir sebentar ke Bank Contoh, bank menengah yang kita kenal di pembahasan sebelumnya. Bank, tokoh, dan angkanya rekaan; kasus Air Canada tadi nyata.

Chatbot nasabah Bank Contoh sudah setahun melayani pertanyaan kartu kredit. Pak Yusuf, direktur teknologi, baru saja menerima pemberitahuan dari vendornya: model di balik *chatbot* akan diperbarui akhir bulan, “lebih pintar dan lebih hemat”. Setahun lalu ia akan langsung menyambutnya. Kini ia tahu, memasang versi baru tanpa diuji ibarat menyajikan masakan setengah matang kepada tamu: yang pertama tahu rasanya justru pelanggan. Kali ini ia meminta tim

menjalankan ulang kumpulan uji dua ratus pertanyaan yang dibangun sejak peluncuran, pada versi baru itu, sebelum pelanggan mana pun menyentuhnya.

Sembilan puluh empat jawaban tetap benar. Enam yang keliru semuanya berkumpul di satu jenis pertanyaan: biaya keterlambatan pembayaran. Versi baru itu, entah mengapa, menjadi lebih murah hati. Ia menyebut denda bisa dihapus berlaku surut bila nasabah “mengajukan keberatan”, suatu kebijakan yang tidak pernah ada.

Enam dari seratus. Pada ribuan percakapan sebulan, itu puluhan janji yang akan ditagih nasabah kepada petugas cabang, dan, seperti dipelajari sebuah maskapai di Kanada, mungkin kepada hakim.

Bu Ratna, direktur keuangan, tidak perlu diyakinkan lama-lama; ia menyetujui penundaan hari itu juga. Pak Yusuf menunda pembaruan untuk jenis pertanyaan itu, mengalihkannya ke petugas dengan AI sebagai pendapat kedua, dan meminta vendor menjelaskan apa yang berubah. Jenis pertanyaan lain tetap berjalan dengan versi baru. Keputusan itu memakan dua hari dan tidak masuk berita. Itulah wujud peta yang diperbarui.

Berapa mahal memelihara peta seperti itu? Jauh lebih murah daripada kesannya. Seratus kasus dengan jawaban benar disusun sekali, dalam hitungan hari, oleh orang yang memang paling tahu pekerjaannya. Menjalankannya ulang pada versi model baru memakan hitungan jam, dan sebagian besar bisa diotomasi. Yang perlu dijaga hanyalah kebiasaan menambah kasus baru setiap kali ada kesalahan yang benar-benar terjadi, dan satu nama yang bertanggung jawab agar kumpulan itu tidak usang. Bandingkan dengan biaya satu jenis kesalahan yang berulang di ribuan percakapan sebelum ada yang menyadarinya.

Batas yang bergeser juga berarti peta itu tidak hanya memberi kabar buruk. Tugas yang setahun lalu masuk kolom “petugas menjawab lebih dulu” bisa saja, pada pengujian berikutnya, sudah layak naik kelas: boleh dijawab AI lebih dulu dengan pemeriksaan sampel. Perusahaan yang rajin menguji ulang adalah perusahaan yang paling cepat tahu kapan boleh mempercayakan lebih banyak kepada AI, dengan bukti di tangan. Perusahaan yang tidak menguji hanya punya dua pilihan buruk: tetap takut, atau percaya tanpa dasar.

Sikap yang paling sehat adalah percaya, tetapi terus memeriksa. Percaya pada angka dari kumpulan uji sendiri, bukan pada kesan, bukan pada brosur.

Pembaruan yang Membuat Mundur

Kejadian seperti di Bank Contoh bisa terjadi sungguhan. Model memang bisa mundur di tugas tertentu ketika diperbarui, dan ada contoh terukurnya.

Tiga peneliti, Lingjiao Chen, Matei Zaharia, dan James Zou, membandingkan dua versi GPT-4, versi Maret dan versi Juni 2023, pada sejumlah tugas. Salah satunya sederhana: menentukan apakah sebuah bilangan termasuk bilangan prima. Dalam versi makalah yang terbit di Harvard Data Science Review pada Maret 2024, soalnya berjumlah 1.000, separuh bilangan prima dan separuh bukan. Akurasi GPT-4 turun dari 84,0 persen pada Maret menjadi 51,1 persen pada Juni, hampir setara menebak, sebagian karena versi Juni tidak lagi mengikuti perintah untuk berpikir langkah demi langkah sebelum menjawab.[99]

Temuan ini sempat diperdebatkan. Ketika makalahnya pertama beredar pada Juli 2023, Arvind Narayanan dan Sayash Kapoor menunjukkan bahwa semua soal di versi awal itu bilangan prima; menurut mereka,

hasil makalah itu menunjukkan bahwa cara model menjawab berubah, tetapi belum membuktikan bahwa model menjadi kurang mampu. Namun keduanya juga menulis satu hal yang penting bagi perusahaan: bagi pemakai, model yang berubah cara menjawabnya terasa sama saja dengan model yang menjadi kurang mampu, dan alur kerja yang dibangun di atas model bisa berhenti berfungsi ketika perilakunya bergeser.[100] Ketiga peneliti itu sendiri menyarankan agar perusahaan yang memakai layanan model bahasa dalam alur kerjanya menjalankan pemantauan serupa untuk aplikasinya sendiri.

Pemilik rumah makan yang memakai bumbu racikan pemasok mengenal masalah ini. Pemasok mengganti resep, mungkin demi menekan biaya, dan tidak selalu memberi tahu. Pelanggan yang pertama sadar bahwa rasa sotonya berubah. Pemilik yang cermat mencicipi setiap kiriman baru sebelum masuk panci, bukan karena curiga pada pemasoknya, melainkan karena nama di papan warung adalah namanya sendiri. Model AI yang disewa dari penyedia adalah bumbu racikan itu; kumpulan uji adalah sendok pencicipnya.

Dalam praktik, ada tiga kebiasaan yang perlu dijalankan; yang pertama sudah dilakukan Pak Yusuf tadi. Pertama, jadikan kumpulan uji sebagai patokan: setiap kali penyedia mengumumkan pembaruan, jalankan ulang soal yang sama, dan bandingkan nilainya per jenis kasus, bukan hanya rata-ratanya, karena rata-rata 94 persen menyembunyikan enam kesalahan yang berkumpul di satu tempat. Kedua, uji juga secara berkala walau tidak ada pengumuman; di Bagian Pembahasan "Ketika AI Lepas Kendali" kita akan melihat bahwa model bisa diubah pembuatnya tanpa pemberitahuan. Ketiga, tetapkan sejak awal siapa yang memutuskan bila nilainya turun, dan apa pilihannya: menunda pembaruan, memperketat pemeriksaan manusia, atau memakai jalan mundur yang sudah disiapkan.

Batas itu memang bergeser. Kadang maju, kadang mundur. Yang bisa kita lakukan adalah terus mengukurnya.

Di Dalam dan di Luar Batas

AI yang sama bisa sangat membantu atau justru menyesatkan, tergantung tugasnya.

Situasi	Yang terjadi	Sumber
Konsultan, tugas di dalam batas	12,2% lebih banyak tugas, 25,1% lebih cepat, mutu lebih dari 40% lebih tinggi	Eksperimen acak BCG, 758 konsultan (2023; jurnal 2026)
Konsultan, tugas di luar batas	19 poin persen lebih kecil kemungkinan menjawab benar	Eksperimen yang sama
Dokter dengan AI	Skor 76% vs 74% tanpa AI; selisih tidak bermakna secara statistik. AI sendiri 16 poin lebih tinggi	Uji klinis acak, JAMA Network Open, 2024
Alat riset hukum khusus	Berhalusinasi 17–33%; *chatbot* umum 43% (alat diuji Maret–Mei 2024)	Studi Stanford RegLab, 2025
Alat riset hukum, setahun kemudian	Akurasi 78–81%, *chatbot* umum 80% (alat pertengahan 2025)	Vals AI, Oktober 2025
Pengembang berpengalaman	Merasa 20% lebih cepat, terukur 19% lebih lambat (2025); pembaruan 2026 condong lebih cepat, bukti lemah	Uji acak METR
Tugas berbeda, hasil berbeda	Halusinasi 22–94% pada pertanyaan pengetahuan (26 model); 1,8–5,4% pada tugas meringkas (15 model teratas)	AI Index 2026
Model yang sama, versi berbeda	Soal bilangan prima: 84,0% (Maret 2023) menjadi 51,1% (Juni 2023)	Chen, Zaharia, dan Zou, 2024

Memo Jumat Sore

Pada Jumat sore, setelah lembar aturan pakai ditandatangani, Pak Hendra mengirim memo pendek kepada semua kepala unit PT Contoh Distribusi. Memo itu rekaan; isinya bisa Anda tiru besok pagi.

> Kepada semua kepala unit.
>
> Mulai bulan depan, setiap usulan pemakaian AI untuk sebuah tugas datang bersama tiga hal. Pertama, kumpulan uji: sedikitnya seratus kasus nyata dari unit Anda, dengan jawaban benar yang ditetapkan orang paling senior di unit itu, bukan oleh vendor dan bukan oleh tim TI. Kedua, hasilnya per jenis kasus, bukan satu angka rata-rata. Ketiga, aturan pakai untuk setiap jenis kasus, yang ditentukan oleh dua hal sekaligus: seberapa sering AI benar, dan seberapa mahal bila ia keliru.
>
> Tugas yang nilainya tinggi dan akibat kelirunya ringan boleh diserahkan ke AI, dengan sampel yang diperiksa tiap minggu. Tugas yang akibat kelirunya berat tetap dijawab manusia lebih dulu, dengan AI sebagai pendapat kedua, berapa pun nilai ujinya. Setiap kali vendor memperbarui model, kumpulan uji dijalankan ulang sebelum pelanggan menyentuhnya; yang memutuskan bila nilainya turun adalah kepala unit, bukan vendor.
>
> Satu lagi. Demo yang memukau bukan bukti. Angka dari meja sendiri adalah bukti.
>
> Hendra

Mari kembali sebentar ke 758 konsultan di awal bab ini. Mereka profesional bergaji tinggi dan sangat terampil; para penelitinya menyebut mereka kelompok elite dengan gelar dari universitas papan atas dunia seperti Harvard, Yale, dan Oxford.[81] Namun ketika mendapat tugas di luar batas kemampuan AI, mereka yang memakai AI justru lebih jarang sampai pada jawaban yang benar.

Kalau konsultan terlatih bisa tergelincir, bagaimana dengan karyawan kita yang setiap hari memakai AI di sela pekerjaan yang menumpuk?

AI tidak memberi tahu kita kapan ia sedang berada di luar batasnya. Kita sendiri pun sering tidak sadar, bahkan merasa lebih cepat ketika sebenarnya lebih lambat. Karena itu batasnya harus dipetakan, diukur dengan data sendiri, dan diperbarui setiap kali AI-nya berubah. Yang paling tahu jawaban benar untuk pekerjaan sehari-hari adalah para pemilik proses dan orang-orang senior di unit bisnis; merekalah yang menyusun kumpulan uji. Para pemimpin cukup menolak satu kebiasaan: menyetujui pemakaian AI hanya karena demonya memukau.

AI tidak akan memberi tahu kita di mana batasnya. Tugas kitalah untuk menemukannya, sebelum pelanggan yang menemukannya lebih dulu.

Peta batas itu menjawab satu pertanyaan: di mana AI bisa dipercaya. Bab berikutnya membahas pertanyaan yang lebih tidak nyaman. Ketika AI berada di dalam batas dan hampir selalu benar, apa yang terjadi pada manusia yang bertugas memeriksanya?

Panduan Daftar Pustaka

Halaman-halaman sebelum ini adalah isi bab *Batas Acak Kemampuan AI* (Bab 5), dengan satu halaman Identitas Dokumen yang disisipkan setelah halaman pembuka bab. Halaman-halaman setelah ini adalah Daftar Pustaka bab tersebut, memuat 23 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, 18, 80–100. 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 “Batas Acak Kemampuan AI | 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 5: Batas Acak Kemampuan AI. 23 sumber.

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

18. Lamarre, E., K. Smaje, dan R. Levin. *Rewired: McKinsey's Playbook on How Leading Companies Win with Technology and AI*. Edisi ke-2. John Wiley & Sons, 2026.

80. Feynman, Richard P. "Cargo Cult Science." Pidato wisuda California Institute of Technology, 1974. Dimuat di *Engineering and Science*. https://calteches.library.caltech.edu.

81. Dell'Acqua, Fabrizio, Edward McFowland III, Ethan Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, dan Karim R. Lakhani. "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality." Harvard Business School Working Paper 24-013, 22 September 2023. Juga terbit di *Organization Science* 37, no. 2 (Maret–April 2026), daring 11 Maret 2026.

82. Moravec, Hans. *Mind Children: The Future of Robot and Human Intelligence*. Harvard University Press, 1988.

83. Stanford Institute for Human-Centered Artificial Intelligence. *The 2026 AI Index Report*.

84. Goh, Ethan, dkk. "Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial." *JAMA Network Open* 7, no. 10 (2024): e2440969.

85. Hadhazy, Adam. "Can AI Improve Medical Diagnostic Accuracy?" Stanford HAI News, 28 Oktober 2024.

86. Magesh, Varun, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning, dan Daniel E. Ho. "Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools." *Journal of Empirical Legal Studies* 22 (2025): 216–242.

87. "ChatGPT Phony Legal Filing Case Gets Lawyers a $5,000 Fine." Bloomberg Law, 22 Juni 2023.

88. METR (Joel Becker, Nate Rush, Beth Barnes, David Rein). "Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity." metr.org, 10 Juli 2025. Makalah lengkap di arXiv 2507.09089.

89. METR. "We Are Changing Our Developer Productivity Experiment Design." metr.org, 24 Februari 2026.

90. Mollick, Ethan. "Centaurs and Cyborgs on the Jagged Frontier." *One Useful Thing*, 16 September 2023.

91. Ramakrishnan, Rama. "A Practical Guide to Gaining Value From LLMs." *MIT Sloan Management Review*, 4 November 2024.

92. Yuliastuti, Desy. "Halodoc Luncurkan AIDA, Teknologi AI Pendamping Dokter di Indonesia." Fortune Indonesia, 21 Agustus 2025.

93. Yellow.ai. "Lion Air Group transforms customer service with 90% automation powered by Yellow.ai's dynamic AI agent." Studi kasus. https://yellow.ai/case-study/lion-air.

94. Vals AI. "VLAIR – Legal Research." Laporan industri, diperbarui 14 Oktober 2025. https://vals.ai.

95. LawNext. [Liputan atas laporan VLAIR Legal Research]. Oktober 2025.

96. Xu, Frank F., dkk. “TheAgentCompany: Benchmarking LLM Agents on Consequential Real World Tasks.” arXiv 2412.14161, Desember 2024, revisi September 2025.

97. *Moffatt v. Air Canada*, 2024 BCCRT 149. Civil Resolution Tribunal British Columbia.

98. Law360 Canada. [Liputan atas putusan *Moffatt v. Air Canada*]. 16 Februari 2024.

99. Chen, Lingjiao, Matei Zaharia, dan James Zou. “How Is ChatGPT’s Behavior Changing Over Time?” *Harvard Data Science Review*, 12 Maret 2024.

100. Narayanan, Arvind, dan Sayash Kapoor. “Is GPT-4 Getting Worse Over Time?” *AI Snake Oil*, 19 Juli 2023.