KALA AI MULAI BERTINDAK

"…the purpose put into the machine is the purpose which we really desire…"

"…tujuan yang kita masukkan ke dalam mesin adalah tujuan yang benar-benar kita inginkan…"

— Norbert Wiener, matematikawan, 1960[136]

> **Pertanyaan bab ini:** Ketika AI tidak lagi hanya menjawab, tetapi mengembalikan dana, mengubah data, dan mengirim pesan atas nama perusahaan, seberapa jauh kita boleh melepasnya?
>
> **Jawaban singkat:** Lepaskan sejauh yang sudah terbukti pada pekerjaan perusahaan sendiri, dan lepaskan bertahap. Saat ini, agen AI bisa diandalkan untuk pekerjaan dengan jenis yang terbatas, aturan yang jelas, volume yang besar, serta tindakan yang bisa dibatalkan jika keliru. Di luar pekerjaan semacam itu, kesalahan kecil di satu langkah terbawa ke langkah-langkah berikutnya dan makin besar. Tetapkan kebebasan agen untuk setiap jenis tindakan, bukan untuk agennya sekaligus, dan tuliskan kapan agen harus dihentikan sebelum ia mulai bekerja.
>
> **Alat yang Anda bawa pulang:** enam level otonomi, A0 sampai A5; daftar kondisi berhenti; dan lembar Uji Dompet yang diisi sebelum agen diberi akses ke sistem dan uang perusahaan.

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 Ketiga: Berkenalan dengan AI Agent
BAB	Bab 8: Kala AI Mulai Bertindak
HALAMAN	191–218 (nomor halaman cetak buku)
ISI BERKAS	Isi bab, Panduan Daftar Pustaka, dan Daftar Pustaka bab ini

SELASA, 27 FEBRUARI 2024. Klarna, perusahaan teknologi keuangan asal Swedia yang terkenal dengan layanan beli sekarang bayar nanti, mengumumkan hasil sebulan pertama asisten AI-nya. Asisten yang dibangun dengan teknologi OpenAI itu tersedia di 23 pasar dan berbicara dalam lebih dari 35 bahasa. Dalam sebulan, ia melayani 2,3 juta percakapan, dua pertiga dari seluruh obrolan layanan pelanggan Klarna. Menurut perusahaan, beban kerjanya setara dengan 700 petugas layanan penuh waktu. Klarna menyebut skor kepuasan pelanggannya setara dengan petugas manusia, dan urusan pelanggan kini selesai dalam waktu kurang dari 2 menit, padahal sebelumnya 11 menit. Perusahaan memperkirakan asisten itu akan menyumbang perbaikan laba sekitar 40 juta dolar AS sepanjang 2024.[137]

Asisten itu tidak hanya menjawab pertanyaan. Menurut Klarna, ia mengurus pengembalian dana dan retur barang, juga urusan yang berkaitan dengan pembayaran. Pelanggan tetap boleh memilih berbicara dengan petugas manusia.[137] Untuk memberi gambaran skala, CEO Klarna, Sebastian Siemiatkowski, menulis di X bahwa layanan pelanggan perusahaannya selama ini ditangani rata-rata 3.000 petugas penuh waktu di perusahaan mitra alih daya.[138] Dengan angka perusahaan itu, hampir seperempat beban kerja para petugas kini dipikul satu sistem.

KAMIS, 8 MEI 2025. Pemimpin yang sama berbicara kepada Bloomberg dengan nada yang jauh berbeda. Menurutnya, biaya ternyata menjadi pertimbangan yang terlalu dominan ketika layanan itu disusun, sehingga, dalam kata-katanya sendiri, “what you end up having is lower quality”.[139]

Dua tanggal itu memuat pelajaran bab ini. AI yang mulai bertindak atas nama perusahaan bisa menghemat sangat banyak. Ia juga bisa pelan-

pelan menurunkan mutu layanan, hal yang tidak tercatat di laporan biaya, sampai pemimpinnya sendiri mengakuinya di depan wartawan.

Bab ini membuka Bagian Ketiga, tentang agen AI. Di Bagian Kedua kita membahas AI yang menjawab, meringkas, dan memberi saran, sementara manusia yang memutuskan. Mulai sekarang kita membahas AI yang mengambil tindakan sendiri atas nama perusahaan.

Dari AI yang Menjawab ke AI yang Bertindak

Sebuah keluarga baru saja mendapat asisten rumah tangga yang sangat pintar.

Pada minggu pertama, sang ibu hanya bertanya kepadanya. Masak apa malam ini? Berapa kira-kira belanjaan yang dibutuhkan? Toko mana yang lebih murah untuk beras dan minyak goreng? Asisten itu menjawab dengan baik. Kalau jawabannya keliru, keluarga itu sendiri yang menanggungnya, karena sang ibu yang memutuskan dan berangkat belanja.

Lalu suatu pagi, karena sibuk, sang ibu memberinya dompet dan kunci motor. Silakan belanja sendiri, bayar sendiri, dan atur sendiri isi kulkas sampai akhir pekan.

Asistennya dan kepintarannya tetap sama, tetapi taruhannya berubah total. Kesalahan kecil yang dulu hanya berupa saran keliru kini bisa berupa uang yang sudah terbelanjakan, atau sekarung beras yang sudah dibeli padahal di gudang masih ada dua.

Itulah lompatan dari AI percakapan ke agen AI. AI percakapan menjawab, meringkas, dan menyusun draf; manusia yang memutuskan dan mengeksekusi. Agen AI melangkah lebih jauh. Ia bisa menjalankan serangkaian langkah lintas berbagai sistem: membaca permintaan

pelanggan, mengecek data pesanan, memproses pengembalian dana, mengirim konfirmasi, lalu mencatat semuanya. Asisten Klarna di awal bab sudah termasuk agen AI dalam arti ini.

Nilai terbesar AI memang baru muncul ketika ia benar-benar menyelesaikan pekerjaan, bukan hanya menyarankannya. Kisah C.H. Robinson dalam cerita sebelumnya adalah contohnya: AI yang mengeksekusi pekerjaan dalam alur kerja yang sudah dibenahi ikut mendorong margin, menurut laporan perusahaan itu sendiri. Karena itu, saran saya, mulailah dari lingkaran kerja yang kecil tetapi utuh: satu jenis interaksi layanan pelanggan, satu jenis peninjauan kontrak, atau satu pengecekan persediaan. Utuh berarti agen membaca permintaan, mengeksekusi tindakannya, mencatat apa yang ia lakukan, dan hasilnya diukur.[4] AI yang berhenti pada tahap memberi saran baru setengah jalan.

Namun memberi dompet kepada asisten baru tetap butuh pertimbangan. Seberapa sering ia salah hitung? Berapa banyak uang yang boleh ia pegang? Belanjaan apa yang tetap harus seizin pemilik rumah? Kapan pula dompet itu diminta kembali? Untuk agen AI, keempat pertanyaan itu menjadi: seberapa sering agen keliru, berapa uang yang boleh ia pegang, tindakan apa yang harus menunggu izin, dan kapan agen dihentikan.

Kesalahan yang Beranak-pinak

Seberapa andal agen AI hari ini? Jawabannya bergantung pada siapa yang mengukur, dan bagaimana mengukurnya.

Pada Mei 2025, tim riset Salesforce menguji agen-agen terkemuka pada tugas bisnis sehari-hari di sistem pengelolaan pelanggan. Bila urusannya bisa selesai dalam satu giliran, agen berhasil sekitar 58 persen. Bila agen

harus bolak-balik bertanya jawab dengan pengguna, seperti percakapan layanan pelanggan sungguhan, keberhasilannya turun ke 35 persen.[140] Para peneliti juga mencatat bahwa tanpa perintah khusus, kesadaran agen untuk menjaga informasi rahasia nyaris nol. Temuan itu akan kita temui lagi di Bagian Pembahasan "Ketika AI Lepas Kendali" dan Bagian Pembahasan "Keamanan Dimulai dari Pembatasan Akses".

Hasil lain sejalan. Di Bagian Pembahasan "Desain Ulang: Benahi Dapurnya Dulu" kita berkenalan dengan τ-bench, uji yang meminta agen melayani pelanggan sambil mematuhi aturan perusahaan. Menurut laporan AI Index 2026 dari Stanford, tidak ada model yang melampaui 71 persen di uji ini.[83] Artinya, agen terbaik pun masih gagal hampir satu dari tiga percobaan. Laporan yang sama mengingatkan bahwa tolok ukur yang banyak dipakai pun tidak bebas cacat; tingkat kesalahannya bisa mencapai 42 persen pada salah satu uji matematika yang populer.[83] Karena itu, skor di papan peringkat belum cukup untuk memutuskan pembelian, apalagi angka di brosur vendor.

Memang ada angka yang jauh lebih tinggi. Pada Agustus 2026, sebuah model besar asal Tiongkok tercatat meraih 86,1 persen di OSWorld-Verified, uji yang meminta agen mengoperasikan komputer seperti manusia. Untuk uji yang sama, angka tertinggi yang dicatat AI Index 66,3 persen.[141] Angka pertama dilaporkan pembuatnya sendiri, sedangkan angka kedua dicatat tim Stanford penyusun laporan itu.

Mengapa agen begitu sulit diandalkan? Salah satu sebabnya bisa dihitung dengan peluang sederhana. Ambil agen yang benar 97 persen di setiap langkah. Bila ia harus melewati 34 langkah berturut-turut tanpa pemeriksaan, peluangnya menuntaskan seluruh rangkaian dengan benar hanya sekitar 36 persen. Naikkan menjadi 99 persen per langkah, peluangnya baru sekitar 71 persen. Pesannya jelas: makin panjang

rangkaian langkah yang dikerjakan agen tanpa pemeriksaan, makin besar peluang ada yang meleset di tengah jalan.

Anak-anak mengenal gejala ini dari permainan pesan berantai. Kalimat yang dibisikkan dari satu anak ke anak berikutnya hampir selalu berubah ketika sampai di ujung barisan. Tidak ada yang berniat jahat. Kesalahan kecil di setiap tahap ikut terbawa, lalu bertambah. Yang tiba di ujung barisan sering kali terdengar masuk akal, dan justru karena itu tidak ada yang curiga.

Pelajarannya praktis. Alur kerja yang panjang dipecah menjadi bagian-bagian pendek. Di antara bagian-bagian itu dipasang pemeriksaan, entah oleh manusia, entah oleh aturan otomatis yang sederhana. Otonomi agen dibatasi pada rangkaian yang pendek, dan diperpanjang hanya setelah terbukti.

Ada pula soal biaya. Dalam pengalaman Bain, versi pertama alur kerja yang dijalankan agen sering berbiaya jauh di atas perkiraan; sebuah alur kerja baru bisa memakan biaya sekitar sepuluh kali lipat dari yang seharusnya, dan sebagian besar kelebihan itu baru bisa dipangkas setelah tim memahami hitungannya.[142] Agen yang mengerjakan lebih banyak langkah menghabiskan lebih banyak token, dan agen yang berputar-putar mengulang langkah menghabiskannya tanpa hasil. Batas biaya (plafon) untuk setiap agen yang dibahas sebelumnya berlaku penuh di sini.

Semua angka tadi berguna terutama di satu tempat: meja tempat proposal vendor dibaca. Brosur agen AI hampir selalu memuat satu angka keberhasilan yang besar. Saran saya, sebelum angka itu dipercaya, ajukan lima pertanyaan berikut. Empat pertanyaan menyangkut keandalan, dan satu, pertanyaan keempat, menyangkut biaya, yang dasar hitungannya sudah dibahas sebelumnya.

Pertama, uji itu dilaporkan siapa? Di uji OSWorld tadi, angka yang dilaporkan pembuat modelnya sendiri sekitar dua puluh poin lebih tinggi daripada angka tertinggi yang dicatat AI Index.

Kedua, satu giliran atau bolak-balik? Agen yang hebat menjawab satu pertanyaan bisa tersandung ketika harus bertanya balik, menunggu jawaban, lalu melanjutkan. Percakapan layanan pelanggan hampir selalu bolak-balik.

Ketiga, berapa langkah, dan diperiksa di mana? Minta vendor menggambar alur kerja agennya langkah demi langkah, lalu tunjukkan di titik mana kesalahan akan tertangkap sebelum sampai ke pelanggan.

Keempat, berapa biaya per tugas yang benar-benar selesai? Bukan biaya per token atau per lisensi, melainkan biaya untuk satu urusan pelanggan yang tuntas, termasuk waktu manusia yang memeriksanya.

Kelima, ke mana kasus yang gagal pergi? Agen yang tidak punya jalan keluar ke manusia akan terus mencoba, dan setiap percobaan menambah biaya serta menambah kekesalan pelanggan.

Vendor yang yakin pada produknya semestinya tidak keberatan ditanya seperti ini, karena jawabannya membedakan mereka dari pesaing yang hanya punya demo. Bila vendor tersinggung atau menolak menjawab, perlakukan sikap itu sebagai informasi: belum ada bukti bahwa agennya andal untuk pekerjaan Anda. Membeli agen hanya berbekal demo ibarat membeli kucing dalam karung: mutunya baru ketahuan setelah dibayar.

Berapa Lama Agen Boleh Bekerja Sendiri?

Di banyak kantor, ada satu pekerjaan yang selalu dikerjakan orang yang sama di akhir bulan: rekonsiliasi rekening bank. Seorang akuntan senior biasanya butuh setengah hari untuk satu rekening. Ia menarik mutasi,

mencocokkan transaksi dengan buku besar, menandai selisih, menelusuri penyebabnya, lalu menyusun ringkasan. Bila pekerjaan setengah hari itu diserahkan kepada agen sekaligus, dari langkah pertama sampai ringkasan, berapa besar peluang hasilnya benar?

Hitungan langkah tadi memberi satu jawaban. Ada cara lain untuk melihat batas agen: memakai waktu.

METR, lembaga riset nirlaba yang sudah kita temui di Bagian Pembahasan "Batas Acak Kemampuan AI", mengukur rentang waktu tugas. Caranya begini. Para peneliti mengukur berapa lama orang yang ahli di bidangnya menyelesaikan serangkaian tugas perangkat lunak, lalu menguji model AI pada tugas yang sama. Hasilnya dinyatakan dalam menit kerja manusia. Menurut versi makalah yang diperbarui pada Februari 2026, model terdepan seperti o3 punya rentang 50 persen sekitar 110 menit. Artinya, untuk tugas yang biasanya dikerjakan seorang ahli dalam sekitar 110 menit, model itu berhasil kira-kira separuh dari percobaannya. Rentang itu berlipat dua kira-kira setiap tujuh bulan sejak 2019, dan mungkin melaju lebih cepat sejak 2024. Bila hasil ini berlaku untuk pekerjaan perangkat lunak sungguhan, para peneliti memperkirakan bahwa dalam lima tahun AI akan sanggup mengotomasi banyak tugas yang kini memakan waktu sebulan bagi manusia.[143]

Bagi perusahaan, ada angka yang lebih penting. Rentang 80 persen, yaitu batas tugas yang bisa diselesaikan dengan peluang berhasil 80 persen, empat sampai enam kali lebih pendek. Bila rentang 50 persen sekitar 110 menit, rentang 80 persennya kira-kira 18 sampai 28 menit. Para peneliti juga menemukan bahwa kinerja AI jauh lebih rendah pada tugas yang kurang terstruktur, yang mereka sebut “berantakan”.[143] Banyak pekerjaan kantor justru seperti itu. Untuk banyak proses, peluang berhasil 80 persen pun masih jauh dari cukup.

Lalu apa yang bisa dilakukan? Saran saya ada empat.

Pertama, ukur durasi tugas berdasarkan waktu yang dibutuhkan manusia untuk mengerjakannya. Berapa menit yang biasanya dibutuhkan karyawan berpengalaman untuk menyelesaikan satu langkah? Langkah yang memakan waktu lama perlu dipertimbangkan untuk dipecah.

Kedua, pasang titik periksa sebelum agen bekerja terlalu lama tanpa diperiksa. Jarak antartitik periksa sebaiknya jauh lebih pendek daripada panjang tugas yang masih bisa dikerjakan agen dengan andal, misalnya rentang 80 persen tadi. Bila agen menangani rangkaian kerja yang panjang, sisipkan pemeriksaan sebelum hasil satu tahap dipakai untuk tahap berikutnya. Pemeriksaan itu tidak selalu harus dikerjakan manusia. Sebagian bisa berupa aturan otomatis, misalnya mencocokkan total saldo dengan buku besar sebelum ringkasan disusun.

Ketiga, jangan mengandalkan angka dari luar. Rentang waktu METR diukur pada tugas tertentu. Batas yang sebenarnya untuk pekerjaan Anda hanya bisa diketahui lewat kumpulan uji sendiri, dan uji itu diulang setiap kali modelnya diganti.

Keempat, tanyakan panjang tugas kepada vendor. Ketika sebuah agen ditawarkan, tanyakan tugas sepanjang apa yang diuji, dan dengan peluang berhasil berapa. Angka keberhasilan tanpa keterangan panjang tugasnya belum cukup untuk menilai agennya.

Angka 110 menit itu sendiri adalah potret saat makalah ditulis, bukan angka tetap. Bila trennya berlanjut, model yang lebih baru akan punya rentang yang lebih panjang. Yang tidak berubah adalah cara memakainya: rentang yang aman untuk perusahaan selalu jauh lebih pendek daripada rentang yang diumumkan.

Setahun Kemudian di Klarna

Dari sisi biaya, angka Klarna tetap mengesankan. Dalam laporan kuartal pertama 2025, Klarna menyatakan biaya layanan pelanggan per transaksi turun 40 persen dalam dua tahun, dari 32 sen dolar AS pada kuartal pertama 2023 menjadi 19 sen pada kuartal pertama 2025.

Biaya layanan pelanggan dan operasinya 51 juta dolar AS pada kuartal pertama 2025, turun dari 57 juta pada periode yang sama tahun sebelumnya. Menurut laporan itu, kepuasan pelanggan tetap stabil.[144]

Namun dua minggu sebelum angka itu diberitakan, dalam wawancara dengan Bloomberg itu, Siemiatkowski mengakui bahwa upaya mengejar penghematan biaya telah menurunkan mutu. Klarna lalu merekrut lagi tenaga layanan pelanggan manusia, supaya pelanggan selalu punya pilihan berbicara dengan manusia. Perusahaan itu menguji model kerja yang disebutnya mirip pola Uber: tenaga layanan bisa masuk ke sistem dan bekerja dari jarak jauh. Tujuannya, kelak menggantikan beberapa ribu petugas yang selama ini dikontrak dari perusahaan alih daya.[139] Klarna menjanjikan bayaran yang bersaing dan keleluasaan penuh untuk menarik orang-orang terbaik, dan menurut Siemiatkowski, menanam modal pada mutu layanan manusia adalah arah masa depan perusahaannya.[145]

Kisah Klarna sering diceritakan sebagai kisah kegagalan AI. Saya membacanya berbeda. Asisten AI-nya tetap bekerja, angka biayanya tetap turun, dan perusahaan itu tidak mencabut asistennya. Yang berubah adalah keyakinan bahwa angka biaya sudah cukup untuk menilai layanan. Biaya per transaksi turun, tetapi menurut CEO-nya sendiri, mutu ikut turun. Dari kisah ini saya menarik satu prinsip: mutu adalah batasan yang tidak boleh dilanggar, bukan sekadar satu ukuran tambahan di samping biaya. Agen yang menghemat biaya dengan

menurunkan mutu sebenarnya sedang memindahkan biaya ke tempat lain: ke pelanggan yang pergi diam-diam, ke keluhan yang berulang, atau ke reputasi.

Klarna tidak sendirian. Pada Agustus 2024, Alibaba menjalankan uji acak tentang agen AI di layanan pelanggan Taobao, dan hasilnya terbit pada Mei 2026. Agen AI memang memperpendek percakapan. Namun penilaian pelanggan untuk percakapan yang bisa ditangani AI justru turun cukup besar. Ketika percakapan dialihkan ke petugas manusia, campur tangan manusia berhasil menjaga mutu layanan bila pengalihan dipicu kendala teknis. Bila pengalihan dipicu pelanggan yang sedang emosi, campur tangan itu kurang berhasil; para peneliti mencatat petugas mengirim lebih sedikit pesan pada percakapan semacam itu.[146] Studi ini akan kita buka lebih rinci di Bagian Pembahasan "Mempekerjakan Agen AI di Perusahaan Anda".

Ketika AI menangani pertanyaan yang mudah, yang tersisa untuk manusia adalah kasus yang sulit dan pelanggan yang sudah kesal. Pekerjaan petugas tidak otomatis menjadi lebih ringan; malah mungkin lebih berat, dengan imbalan dan pelatihan yang belum menyesuaikan. Perusahaan yang memasang agen di layanan pelanggan perlu merancang ulang pekerjaan manusianya juga, bukan hanya menghitung berapa orang yang tidak lagi dibutuhkan.

Lalu seberapa banyak perusahaan yang benar-benar sudah memasang agen? Menurut survei Gartner terhadap para CIO dan eksekutif teknologi pada 2026, baru 17 persen organisasi yang sejauh ini memasang agen AI, sementara lebih dari 60 persen berharap memasangnya dalam dua tahun ke depan.[147] Sebagian besar perusahaan, dengan kata lain, masih berdiri di tempat Klarna berdiri

pada awal 2024. Mereka punya kesempatan belajar dari babak kedua sebelum mengalaminya sendiri.

Otonomi Secukupnya

Kalau agen AI belum sepenuhnya andal, apakah perusahaan sebaiknya menunggu? Tidak juga.

Saran saya, pegang satu prinsip: otonomi secukupnya. Agen AI sudah bisa menghemat biaya pada pekerjaan dengan jenis yang terbatas, aturan yang jelas, volume yang besar, serta tindakan yang bisa dibatalkan jika keliru, misalnya mencocokkan pembayaran dengan faktur. Di luar pekerjaan semacam itu, keandalan langkah demi langkahnya masih rendah, biayanya tinggi dan naik-turun, dan salah satu celah keamanannya, yang dikenal sebagai *prompt injection* atau penyelundupan perintah, yaitu perintah yang diselipkan orang lain ke dalam pesan atau dokumen yang dibaca agen, belum terpecahkan. Kit akan membahasnya di pembahasan berikutnya. Makin besar otonomi yang diberikan, makin berat pengaman yang wajib dipasang, dan pengaman itu harus berada di luar model AI-nya, bukan hanya berupa perintah agar AI berhati-hati.

Kunci prinsip ini ada pada satu gagasan: otonomi ditetapkan per jenis tindakan, bukan per agen. Satu agen yang sama bisa diberi kebebasan berbeda untuk tindakan yang berbeda. Agen boleh memperbarui status tiket keluhan sendiri. Agen boleh menyiapkan *email* untuk pelanggan, tetapi *email* itu harus disetujui manusia sebelum terkirim. Untuk pembayaran, agen hanya boleh memberi rekomendasi; manusia yang mengeksekusi.

Untuk itu saya menyusun enam level otonomi, dari A0 sampai A5:

Level	Yang boleh dilakukan agen	Peran manusia
A0 Memberi informasi	Menjawab, meringkas, menganalisis tanpa mengubah apa pun	Memakai informasinya
A1 Merekomendasikan	Menyusun draf dan rekomendasi	Mengeksekusi sendiri
A2 Menyiapkan	Menyiapkan tindakan; eksekusi menunggu persetujuan setiap kali	Menyetujui setiap tindakan; di atas batas nominal, dua penyetuju
A3 Mengeksekusi yang kecil	Tindakan kecil yang bisa dibatalkan, dengan batas per tindakan dan per hari	Turun tangan bila batas terlampaui
A4 Menjalankan alur kerja	Rangkaian langkah dalam batas kebijakan yang tetap	Menangani pengecualian; memantau saat itu juga
A5 Otonomi tinggi	Menentukan sendiri subtujuan dan alatnya	Tidak saya rekomendasikan untuk perusahaan pada 2026

Ada dua aturan yang harus didahulukan daripada ketentuan dalam tabel itu. Pertama, tindakan kritis paling tinggi hanya boleh sampai level A2 dan wajib disetujui dua orang. Kedua, keputusan yang di Bagian Pembahasan "Bias Otomasi: Sekadar Tukang Stempel" masuk jenis “manusia dulu”, misalnya kredit, klaim, dan rekrutmen, membatasi agen paling tinggi di level A1, yaitu memberi rekomendasi.

Bagaimana menentukan level yang tepat untuk sebuah tindakan? Saya menyarankan enam pertimbangan: dampaknya, bisa tidaknya dibatalkan, seberapa tidak pasti masukannya, kepekaan datanya, besar uang yang terlibat, dan risiko hukumnya.

Ambil satu contoh: agen layanan pelanggan sebuah perusahaan distribusi yang diminta pelanggan mengubah alamat pengiriman.

Dampaknya sedang, karena barang bisa sampai ke tempat yang salah. Bisa dibatalkan, selama barang belum dikirim. Masukannya tidak pasti, karena permintaan datang lewat pesan singkat yang bisa saja dikirim orang lain. Datanya peka, karena alamat adalah data pribadi. Uang yang terlibat tidak langsung, tetapi satu kiriman bernilai jutaan rupiah bisa hilang. Risiko hukumnya kecil. Dari enam pertimbangan itu, tindakan ini wajar ditempatkan di level A2: agen menyiapkan perubahan, lalu manusia atau pelanggan sendiri, lewat kode konfirmasi, menyetujuinya. Mengubah status pesanan dari “dikemas” menjadi “siap kirim”, sebaliknya, cukup di level A3. Satu agen, dua tindakan, dua level otonomi.

Aturan paling tegas berlaku untuk tindakan kritis, yaitu tindakan yang tidak bisa dibatalkan dan dampaknya besar. Semua perpindahan dana masuk golongan ini, kecuali pengembalian dana kecil ke rekening asal. Untuk tindakan kritis, agen paling jauh hanya boleh menyiapkan, lalu menunggu persetujuan. Wajib ada persetujuan dua orang. Bila ragu, jawabannya tolak. Untuk tugas rutin yang risikonya rendah, pemeriksaan oleh manusia memang biaya, dan biaya itu harus sepadan dengan kesalahan yang dicegahnya. Untuk tindakan kritis, pemeriksaan itu wajib, berapa pun biayanya.

Kapan sebenarnya perusahaan butuh agen? Syaratnya cukup ketat. Pekerjaannya terdiri atas banyak langkah lintas sistem, volumenya tinggi, tindakannya bisa dibatalkan, datanya bersih, hasilnya bisa diukur dalam hitungan menit, dan ada jalur untuk meneruskan kasus ke manusia. Bila pekerjaannya hanya satu langkah, atau aturannya sudah jelas dan pasti, aturan otomatis biasa sering sudah cukup. Ini sejalan dengan Lima Level Teknologi di Bagian Pembahasan "Desain Ulang: Benahi Dapurnya Dulu": pilih level teknologi paling rendah yang sudah cukup menyelesaikan masalah.

Saran saya, tulis kondisi berhenti sebelum agen mulai bekerja, bukan sesudah ada masalah. Hentikan atau turunkan tingkat otonominya bila biaya agen untuk satu tugas yang selesai, ditambah biaya karyawan yang memeriksanya, sudah lebih mahal daripada bila tugas itu dikerjakan karyawan sendiri; bila porsi tugas yang berhasil diselesaikan agen tidak kunjung cukup tinggi untuk membuatnya lebih murah daripada karyawan; bila terjadi satu insiden berat; bila mutunya turun di bawah ambang yang sudah disepakati; atau bila modelnya diganti tanpa uji ulang. Kondisi berhenti yang ditulis hitam di atas putih, artinya tertulis resmi sebelum agen mulai bekerja, jauh lebih mudah ditegakkan daripada yang diperdebatkan setelah uang telanjur keluar.

Memberi agen AI otonomi mirip memberi karyawan baru kewenangan tanda tangan. Tidak ada perusahaan yang memberi staf barunya hak menyetujui pembayaran berapa pun pada hari pertama. Kewenangan diberikan bertahap, sesuai bukti, dan bisa dicabut kapan saja. Agen layak diperlakukan sama.

Mengapa Tim Agen Sering Gagal?

Sebagian vendor menawarkan bukan satu agen, melainkan tim agen: satu agen merencanakan, satu menulis, satu memeriksa, satu lagi berkoordinasi. Rancangannya mirip tim kerja manusia, tetapi dalam pemakaian sungguhan hasilnya sering mengecewakan.

Sekelompok peneliti dari UC Berkeley mencoba memahami sebabnya. Mereka menganalisis 1.642 catatan proses kerja dari tujuh kerangka perangkat lunak yang populer untuk membangun tim agen. Tingkat kegagalan pada tujuh sistem sumber terbuka itu berkisar 41 sampai 86,7 persen, dan keuntungan sistem banyak agen sering kali minim dibanding sistem dengan satu agen.[148]

Dari catatan-catatan itu, mereka menyusun daftar 14 jenis kegagalan dalam tiga kelompok. Kelompok pertama adalah masalah rancangan sistem, yang menurut para peneliti berasal dari keputusan rancangan serta perintah yang buruk atau ambigu. Kelompok kedua adalah ketidakselarasan antaragen, yang muncul ketika arus informasi penting di antara agen terputus saat mereka berinteraksi. Kelompok ketiga adalah verifikasi tugas: pemeriksaan yang tidak menemukan atau tidak membetulkan kesalahan, atau pekerjaan yang dihentikan sebelum waktunya.[148]

Bila persentase tiap jenis kegagalan dijumlahkan per kelompok, menurut hitungan saya porsinya kira-kira 44, 32, dan 24 persen.[148] Jenis kegagalan yang paling sering dicatat adalah mengulang langkah, 15,7 persen: agen berputar-putar mengulang langkah tanpa menemukan solusi yang dibutuhkan, sambil terus menghabiskan biaya, seperti yang kita singgung di Bagian Pembahasan "Langganan AI Dibayar Dolar, tapi Gaji Rupiah". Menurut hitungan tadi, lebih dari dua pertiga kegagalan berkaitan dengan rancangan sistem dan pemeriksaan hasil, bukan semata soal kecerdasan modelnya. Kabar baiknya, kedua hal itu lebih banyak berada di tangan perusahaan yang memasang sistemnya.

Berikut ringkasannya. Porsi di kolom tengah diambil dari makalah; langkah pencegahan di kolom kanan adalah saran saya.

Kelompok kegagalan	**Contoh jenis kegagalan dan porsinya**	**Langkah pencegahan**
Rancangan sistem	Mengulang langkah (15,7%); tidak tahu kapan berhenti (12,4%); tidak mematuhi spesifikasi tugas (11,8%)	Spesifikasi tugas dan kondisi selesai yang tertulis jelas
Ketidakselarasan antaragen	Penalaran tidak sejalan dengan tindakan (13,2%); keluar dari jalur tugas (7,4%); tidak meminta klarifikasi (6,8%)	Satu penanggung jawab koordinasi; agen wajib bertanya bila ragu
Verifikasi tugas	Verifikasi keliru (9,1%); tidak ada atau tidak tuntas (8,2%); berhenti sebelum waktunya (6,2%)	Pemeriksaan berlapis, termasuk terhadap tujuan awal

Bisakah diperbaiki? Sebagian. Para peneliti mencoba dua perbaikan sederhana. Memperjelas spesifikasi dan alur kerja para agen menaikkan keberhasilan satu sistem sebesar 9,4 persen. Menambahkan satu langkah pemeriksaan terhadap tujuan tugas secara keseluruhan memberi perbaikan 15,6 persen pada uji lainnya. Namun para peneliti sendiri mengingatkan bahwa tidak semua jenis kegagalan teratasi, dan tingkat penyelesaian tugas tetap rendah.[148]

Saran saya, jangan menambah agen hanya supaya terlihat canggih. Tulis spesifikasi setiap agen sejelas mungkin: apa tugasnya, apa batasnya, dan kapan pekerjaannya dianggap selesai. Siapkan dua lapis pemeriksaan: pemeriksaan di antara langkah-langkah kerja dan pemeriksaan akhir untuk memastikan apakah hasilnya benar-benar memenuhi tujuan awal. Bila sistemnya sudah berjalan, tiga kelompok tadi bisa dipakai sebagai kategori laporan insiden, supaya tim tahu apakah masalahnya ada di rancangan, di koordinasi, atau di pemeriksaan. Bagian Pembahasan

"Mempekerjakan Agen AI di Perusahaan Anda" akan membahas kapan banyak agen memang layak dipakai.

Sketsa: Agen di Meja Piutang

Perusahaan, tokoh, dan angka dalam sketsa PT Contoh Distribusi ini rekaan; kesamaan dengan perusahaan atau orang sungguhan tidak disengaja. Yang nyata adalah dasarnya: hitungan peluang kesalahan beruntun, temuan para peneliti UC Berkeley tentang kegagalan tim agen, dan beban per jam karyawan yang kita hitung di Bagian Pembahasan "Langganan AI Dibayar Dolar, tapi Gaji Rupiah" dari data upah BPS.

Di pembahasan sebelumnya, Bu Sandra, direktur keuangan, menerima tiga staf baru di bagian rekonsiliasi piutang. Beberapa bulan kemudian, bagian yang sama menerima tawaran dari sebuah vendor. Tawarannya dua agen AI untuk tutup buku piutang setiap bulan. Agen pertama, agen pencatat, menarik mutasi rekening bank, mencocokkan setiap pembayaran pelanggan dengan fakturnya, lalu menandai faktur yang sudah lunas. Agen kedua, agen penagih, menyusun pengingat untuk pelanggan yang belum membayar. Bagian rekonsiliasi mengolah sekitar 1.800 pembayaran pelanggan setiap bulan, dan tutup buku biasanya baru selesai pada hari kerja ketujuh.

Demo vendor berjalan mulus. Semua faktur contoh tercocokkan dalam hitungan menit. Pak Hendra, direktur utama, sudah setengah yakin. Bu Sandra meminta satu hal sebelum memutuskan: “Uji dulu pada kasus kita sendiri.”

Kepala bagian piutang menyiapkan 50 kasus dari arsip yang jawabannya sudah diketahui. Isinya kasus-kasus yang membuat staf baru tersandung: satu transfer untuk beberapa faktur, potongan retur yang belum

dibukukan, pelanggan yang salah mentransfer ke rekening perusahaan lain dalam grup, dan pembayaran dengan jumlah yang tidak bulat.

Uji pertama menjalankan kedua agen seperti dalam demo: rangkaian penuh, dari mutasi bank sampai pengingat pelanggan, tanpa titik periksa di tengah. Hasilnya, 31 kasus benar dari ujung ke ujung, 19 keliru. Sebagian besar kekeliruan muncul pada pembayaran gabungan, persis di tempat staf baru juga paling sering keliru. Namun ada tiga kasus yang membuat rapat terdiam. Tiga faktur sudah dibayar, uangnya sudah masuk, tetapi tidak ditandai lunas oleh agen mana pun. Agen pencatat menganggap penandaan lunas adalah langkah terakhir agen penagih, dan agen penagih menganggapnya bagian dari pencocokan. Akibatnya, agen penagih menyusun pengingat kepada tiga pelanggan yang sudah membayar.

“Kalau pengingat itu terkirim,” kata Bu Sandra, “yang marah bukan hanya pelanggannya. Tim penjualan Pak Dedi yang ditelepon duluan.”

Bentuk kegagalan itu persis yang dicatat para peneliti UC Berkeley: tidak jelas siapa memegang langkah yang mana. Vendor diminta mengubah rancangannya. Alur kerja dipecah menjadi empat tahap: menarik mutasi, mencocokkan, menandai selisih, dan menyusun pengingat. Di antara tahap kedua dan ketiga dipasang aturan otomatis yang mencocokkan total saldo dengan buku besar; bila tidak cocok, kasus berhenti dan diteruskan ke staf. Penanda “lunas” hanya boleh diberikan oleh satu pihak, yaitu agen pencatat, dan agen penagih tidak boleh menyusun pengingat untuk faktur yang belum diperiksa tahap ketiga.

Uji kedua memakai 50 kasus yang sama. Hasilnya, 44 kasus ditangani dengan benar, 5 kasus dihentikan oleh aturan otomatis dan diteruskan ke staf, dan 1 kasus lolos meski keliru, yaitu sebuah selisih kecil yang

tercatat sebagai potongan. Tidak sempurna, tetapi kesalahan yang tersisa kini kecil dan bisa ditelusuri.

Lima kasus yang dihentikan aturan otomatis diserahkan kepada staf. Bu Sandra sengaja memberikannya kepada dua staf baru yang belajar dengan mode tutor. Alasannya sederhana: kasus-kasus itulah yang paling banyak mengajarkan sesuatu. Satu di antaranya ternyata transfer pelanggan yang salah alamat ke rekening perusahaan lain dalam grup, kasus yang tidak mungkin dipecahkan agen karena datanya memang tidak ada di sistem PT Contoh. Staf itu menelepon bagian keuangan perusahaan saudara, dan masalahnya selesai dalam sepuluh menit. Agen mengerjakan yang rutin; manusia mengerjakan yang membutuhkan telepon.

Lalu Bu Sandra menetapkan otonomi per tindakan, memakai enam level otonomi tadi. Menarik mutasi dan mencocokkan pembayaran boleh dieksekusi agen sendiri, karena bisa dibatalkan dan tidak menyentuh pelanggan: level A3. Pengingat kepada pelanggan disiapkan agen, tetapi baru terkirim setelah disetujui staf: level A2. Jurnal penyesuaian dan nota kredit, yang mengubah saldo pelanggan, hanya boleh direkomendasikan agen; staf yang mengeksekusi, dan untuk nilai di atas Rp10 juta harus disetujui Bu Sandra bersama satu orang lain: level A1, dengan persetujuan dua orang.

Terakhir, kondisi berhenti, ditulis sebelum uji coba dimulai. Biaya agen, misalkan, Rp2.000 untuk setiap kasus yang selesai. Pemeriksaan oleh staf rata-rata 3 menit per kasus. Dengan beban per jam Rp26.600 yang kita hitung di Bagian Pembahasan "Langganan AI Dibayar Dolar, tapi Gaji Rupiah", pemeriksaan itu bernilai sekitar Rp1.330, sehingga biaya satu kasus lewat agen sekitar Rp3.330. Bila dikerjakan staf sepenuhnya, satu kasus butuh sekitar 12 menit, atau Rp5.320.[46] Selisihnya sekitar Rp1.990

per kasus. Sedikit demi sedikit, lama-lama menjadi bukit: untuk 1.800 pembayaran, selisih itu menjadi sekitar Rp3,6 juta sebulan.

Bu Sandra menulis empat kondisi berhenti di bawah hitungan itu. Agen dihentikan bila waktu pemeriksaan per kasus melewati sekitar 7,5 menit, karena pada titik itu agen tidak lagi lebih murah daripada staf. Agen dihentikan bila hasil uji ulang bulanan pada 50 kasus arsip turun di bawah 85 dari 100. Agen dihentikan bila terjadi satu insiden berat, misalnya pengingat terkirim kepada pelanggan yang sudah membayar. Agen juga dihentikan bila vendor mengganti modelnya tanpa uji ulang.

Pak Hendra membaca semuanya lalu bertanya satu hal. “Rp3,6 juta sebulan itu benar-benar kita hemat, atau hanya di atas kertas?”

Bu Sandra tersenyum. Pertanyaan yang sama pernah muncul di rapat uji coba bagian administrasi penjualan di awal buku. “Di atas kertas, Pak, sampai kita putuskan jam staf yang terbebas dipakai untuk apa. Usul saya, dipakai untuk menagih piutang yang sudah lewat 60 hari, yang selama ini tidak pernah sempat dikejar.”

Uji coba tiga bulan disetujui sore itu.

Di Mana Agen Benar-benar Bernilai?

Menurut laporan Klarna sendiri, penghematannya nyata: biaya per transaksi turun 40 persen dalam dua tahun, dan asisten AI-nya tetap menangani sebagian besar percakapan. Yang dikoreksi Klarna bukan keputusannya memakai AI, melainkan anggapan bahwa manusia bisa disingkirkan sepenuhnya dari layanan pelanggan.

Banyak perusahaan sebenarnya sudah memilih jalan tengah. Dalam survei Bain terhadap 293 pemimpin teknologi senior pada April 2026, hanya 6 persen yang membayangkan pengembangan perangkat lunak yang sepenuhnya otonom. Sebanyak 41 persen memperkirakan akan memakai model berjenjang menurut risiko.[149] Itu sejalan dengan gagasan otonomi per jenis tindakan.

Kemampuan agen juga terus berkembang. Angka 35 persen untuk percakapan bolak-balik hari ini bisa tampak kuno beberapa tahun lagi, dan rentang waktu tugas terus memanjang. Perusahaan yang menunggu sampai agen sempurna akan tertinggal dari perusahaan yang sudah belajar menjalankannya dengan pengaman yang benar. Posisi saya sederhana: tancap gas, artinya bergerak cepat, pada pekerjaan yang batasnya jelas dan tindakannya bisa dibatalkan jika keliru, sambil mengumpulkan bukti untuk memperluas penggunaannya.

Di Indonesia, level-level otonomi ini sudah bisa dikenali. Dua contoh yang diberitakan terbuka masih berada di level paling bawah. BRI, misalnya, mengembangkan *chatbot* Sabrina dengan teknologi ChatGPT melalui Azure OpenAI Service, sehingga mampu memahami berbagai bahasa daerah.[150] Telkomsel memasang Digital Smart Care berbasis AI sebagai sistem manajemen pengetahuan untuk para petugas layanan pelanggannya.[39] Menurut sumbernya, keduanya menjawab dan membantu; sumber tidak menyebut keduanya mengeksekusi transaksi. Saya menempatkan keduanya di sekitar level A0: memberi informasi kepada nasabah atau kepada petugas.

Lion Air Group sudah selangkah lebih jauh. Menurut studi kasus vendornya, Yellow.ai, agen AI di WhatsApp dan situs web Lion Air menangani pemesanan, penjadwalan ulang, informasi penerbangan, klaim bagasi, kebijakan perjalanan, dan keluhan; pertanyaan yang rumit

diteruskan ke petugas manusia.[93] Penjadwalan ulang adalah tindakan, bukan sekadar jawaban, dan jalur ke petugas manusia sudah disiapkan. Itulah dua ciri yang dibahas bab ini: tindakan yang bisa dibatalkan, dan jalan keluar ke manusia.

Bagi perusahaan Indonesia yang baru mulai, saran saya, mulailah dari pekerjaan yang paling cocok untuk agen: volumenya besar, langkahnya jelas, tindakannya bisa dibatalkan jika keliru, dan hasilnya terukur. Contohnya memilah dan memperbarui status tiket internal, mencocokkan pembayaran dengan faktur seperti di PT Contoh, atau menyiapkan draf balasan pelanggan yang tetap dicek manusia. Jangan mulai dari pembayaran keluar, keputusan kredit, atau hal-hal yang tidak bisa ditarik kembali. Perusahaan menengah tanpa tim TI yang besar tidak perlu membangun agen sendiri untuk mulai. Bila aplikasi akuntansi, sistem tiket, atau layanan pesan yang sudah dipakai menawarkan fitur agen di dalamnya, pertanyaannya sama saja: tindakan apa yang dibolehkan, pada level berapa, dan siapa yang mencabut aksesnya bila ada yang salah. Untuk perusahaan sebesar apa pun, saran saya sama: pastikan ada satu orang yang namanya tertulis sebagai pemegang kunci, yaitu orang yang berwenang mencabut akses agen.

Sketsa: Pengembalian Biaya yang Dibatasi

Bank Contoh, bank menengah yang kita kenal sejak awal cerita, serta tokoh dan angka dalam sketsa ini rekaan; kesamaan dengan pihak nyata tidak disengaja. Yang nyata adalah kasus-kasus yang menjadi dasarnya: asisten Klarna yang mengurus pengembalian dana, uji acak Alibaba tentang pelanggan yang emosi, dan putusan tentang *chatbot* Air Canada yang kita bahas di Bagian Pembahasan "Batas Acak Kemampuan AI".

Dalam cerita itu, Bank Contoh menguji *chatbot* nasabah yang menjawab pertanyaan soal biaya keterlambatan kartu kredit. Kini Pak Yusuf, direktur teknologi, mengusulkan langkah berikutnya. *Chatbot* itu dinaikkan menjadi agen yang bisa langsung mengembalikan biaya keterlambatan ke rekening asal nasabah, bila nasabah memenuhi aturan bank: keterlambatan pertama, dan paling lama tiga hari. Selama ini permohonan semacam itu menumpuk di antrean petugas, padahal hampir semuanya rutin.

Bu Ratna, direktur keuangan, tidak langsung menolak. Ia bertanya satu hal: berapa ongkosnya bila agen keliru?

Tim Pak Yusuf menjawabnya dengan uji. Mereka mengambil 200 percakapan permohonan dari arsip, yang keputusannya sudah diketahui. Agen memberi putusan benar pada 186 percakapan. Pada 11 percakapan, agen menyetujui pengembalian dana kepada nasabah yang sebenarnya tidak berhak menerimanya. Pada 3 percakapan, agen menolak nasabah yang sebenarnya berhak.

Salah satu dari 11 kekeliruan itu dibacakan di rapat. Seorang nasabah menulis bahwa ia terlambat membayar karena sedang dirawat di rumah sakit, dan meminta biaya keterlambatannya dikembalikan. Menurut catatan, keterlambatannya bukan yang pertama. Menurut aturan bank, permohonan itu ditolak; petugas mungkin saja memutuskan lain dengan pertimbangannya sendiri. Agen menyetujui permohonan itu, dengan kalimat yang sangat santun. Tidak ada yang keberatan dengan hasilnya. Yang dipersoalkan adalah bahwa agen mengambil keputusan di luar aturan tanpa memberi tahu siapa pun. Keputusan semacam itu, kata Bu Ratna, boleh diambil, tetapi oleh orang yang namanya tercatat.

Bu Ratna menghitung potensi kerugian dalam skenario terburuk di papan. Misalkan tiga bulan pertama datang sekitar 3.000 permohonan

sebulan. Bila agen langsung diizinkan mengeksekusi sendiri, dengan tingkat kekeliruan seperti di uji, sekitar 165 pengembalian keliru bisa terjadi setiap bulan. Dengan batas Rp150.000 per pengembalian, potensi kerugiannya paling banyak sekitar Rp24,75 juta sebulan. Angkanya tidak menghancurkan bank. Namun bagi Bu Ratna, yang lebih mengganggu adalah nasabah yang berhak tetapi ditolak: 3 dari 200 dalam uji tadi. Mereka tidak akan muncul di laporan keuangan; mereka akan muncul di media sosial.

Rancangan akhirnya memakai tabel level otonomi tadi. Pengembalian ke rekening asal sampai Rp150.000 per tindakan, dengan batas total Rp20 juta per hari, kelak boleh dieksekusi agen sendiri: level A3. Permohonan di atas batas itu, atau permintaan agar dana dikirim ke rekening lain, disiapkan agen tetapi harus disetujui petugas: level A2. Perpindahan dana apa pun selain pengembalian kecil ke rekening asal wajib disetujui dua orang. Nasabah yang menulis dengan nada marah, atau menyebut kata keluhan, langsung diteruskan ke petugas berpengalaman, bukan ke petugas yang paling longgar antreannya. Pelajaran dari Alibaba dipakai di sini: pelanggan yang emosi butuh orang yang paling siap, bukan orang yang kebetulan kosong.

Lalu keputusan yang paling penting: agen tidak langsung naik ke level A3. Selama delapan minggu pertama, setiap pengembalian disiapkan agen tetapi disetujui petugas, sehingga agen bekerja di level A2. Agen baru naik ke level A3 bila pada uji ulang 200 percakapan, agen keliru menyetujui permohonan dari nasabah yang tidak berhak paling banyak 2 kali, atau 1 dari 100.

Pak Yusuf sempat keberatan. Delapan minggu di level A2 berarti petugas masih harus mengklik setuju ribuan kali. Bu Ratna mengingatkan kisah Air Canada di Bagian Pembahasan "Batas Acak Kemampuan AI". Tribunal

di Kanada menolak dalih bahwa *chatbot* adalah entitas terpisah yang bertanggung jawab atas tindakannya sendiri.[97,98] Apa pun yang dilakukan agen Bank Contoh, nama banklah yang tercantum di mutasi rekening nasabah.

Rancangan itu disetujui dengan satu catatan dari Bu Ratna: jumlah putusan yang benar (pada uji pertama 186 dari 200) dilaporkan ke direksi setiap kali agen diuji ulang, bersama jumlah nasabah berhak yang ditolak keliru (pada uji pertama 3). Kalau jumlah nasabah yang ditolak keliru tidak turun, agen tidak naik level, berapa pun penghematannya.

Uji Dompet: Empat Baris Sebelum Agen Bekerja

Kembali ke PT Contoh Distribusi. Seminggu sebelum uji coba agen piutang dimulai, Pak Hendra memanggil Bu Sandra ke ruangannya. Di mejanya ada selembar kertas dengan empat baris kosong. Ia menyebutnya Uji Dompet.

“Sebelum dompetnya kita serahkan,” katanya, “saya ingin keempat baris ini terisi. Kalau tidak bisa diisi dalam satu halaman, berarti kita belum siap.”

Mereka mengisinya berdua. Hasilnya seperti ini.

1. Apa yang ingin dicapai, dan apa yang tidak boleh dikorbankan. Tutup buku piutang selesai pada hari kerja ketiga, bukan ketujuh. Yang tidak boleh dikorbankan: tidak ada faktur yang ditandai lunas tanpa bukti uang masuk, dan tidak ada pelanggan yang menerima pengingat untuk faktur yang sudah dibayar.

2. Tindakan mana yang boleh dikerjakan agen sendiri, mana yang menunggu izin, dan mana yang tidak boleh disentuh. Sendiri: menarik mutasi bank, mencocokkan pembayaran, menandai selisih. Menunggu izin staf: pengingat kepada pelanggan. Tidak boleh dieksekusi agen: jurnal penyesuaian dan nota kredit (agen hanya merekomendasikan, staf yang membukukan), serta apa pun yang mengubah saldo pelanggan atau memindahkan uang.

3. Kapan agen dihentikan. Bila waktu pemeriksaan per kasus melewati sekitar 7,5 menit; bila uji ulang bulanan turun di bawah 85 dari 100; bila satu pengingat terkirim kepada pelanggan yang sudah membayar; atau bila vendor mengganti model tanpa uji ulang.

4. Siapa yang mencabut kunci, dan seberapa cepat. Bu Sandra, atau kepala bagian TI bila Bu Sandra tidak bisa dihubungi. Hari itu juga, tanpa menunggu rapat.

Baris pertama itulah yang paling lama dibaca ulang Pak Hendra. Baris itu adalah peringatan Norbert Wiener di awal bab ini dalam bahasa kantor: bila kita memakai mesin yang bertindak terlalu cepat untuk dihentikan, kita harus benar-benar yakin bahwa tujuan yang kita masukkan ke dalamnya adalah tujuan yang sungguh kita inginkan. Menurut pengakuan CEO-nya sendiri, Klarna terlalu menitikberatkan biaya ketika menyusun layanannya. Biaya memang turun, tetapi mutu ikut turun.

Lembar seperti ini bisa disalin untuk agen mana pun di perusahaan Anda. Saran saya, lembar itu diisi oleh dua orang: pemilik proses yang tahu apa yang tidak boleh dikorbankan, dan orang yang memegang kas atau akses sistem, yang tahu seberapa cepat kunci bisa dicabut. Di perusahaan besar, lembar itu dilampirkan pada setiap usulan agen, di samping Satu Lembar dari Bagian Pembahasan "Penghematan di Atas Kertas". Di perusahaan kecil, lembar itu cukup ditempel di dinding ruang keuangan.

Bentuknya tidak penting. Yang penting, keempat baris itu ditulis sebelum agen mulai bekerja, ketika semua orang masih tenang dan belum ada uang yang telanjur keluar. Empat barisnya sederhana, tetapi baris ketiga dan keempat, kapan berhenti dan siapa yang mencabut kunci, layak diperiksa dua kali, karena keduanya baru terasa penting ketika keadaan sudah tidak tenang.

Pertanyaannya, apa yang terjadi bila kedua baris itu memang tidak pernah diisi, dan agennya sudah telanjur memegang kunci? Bab berikutnya berisi kisah-kisah nyata tentang itu, dari basis data yang terhapus sampai agen yang keluar dari ruang uji.

Panduan Daftar Pustaka

Halaman-halaman sebelum ini adalah isi bab *Kala AI Mulai Bertindak* (Bab 8), 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, 39, 46, 83, 93, 97, 98, 136–150. 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 “Kala AI Mulai Bertindak | 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 8: Kala AI Mulai Bertindak. 22 sumber.

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

39. IDC. “IDC Announces the 2024 Future Enterprise Awards Winners for Indonesia.” Siaran pers. https://idc.com.

46. Badan Pusat Statistik. “Tingkat Pengangguran Terbuka (TPT) sebesar 4,68 persen. Rata-rata upah buruh sebesar 3,29 juta rupiah.” Siaran pers, 5 Mei 2026. https://www.bps.go.id/id/pressrelease/2026/05/05/2574/tingkat-pengangguran-terbuka--tpt--sebesar-4-68-persen--rata-rata-upah-buruh-sebesar-3-29-juta-rupiah.html.

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

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.

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.

136. Wiener, Norbert. “Some Moral and Technical Consequences of Automation.” *Science* 131, no. 3410 (6 Mei 1960): 1355–1358. https://cs.umd.edu.

137. Klarna. “Klarna AI Assistant Handles Two-Thirds of Customer Service Chats in Its First Month.” Siaran pers, 27 Februari 2024. https://klarna.com/international/press.

138. “Klarna Asked to Be OpenAI’s Guinea Pig. Now Its Chatbot Does the Work of 700 Employees.” HR Grapevine, 29 Februari 2024.

139. “Klarna Focuses on Human-Powered Customer Service.” PYMNTS, 8 Mei 2025.

140. Salesforce AI Research. “CRMArena-Pro: Holistic Assessment of LLM Agents Across Diverse Business Scenarios and Interactions.” arXiv:2505.18878, 24 Mei 2025. Pracetak.

141. Steel.dev. “OSWorld Leaderboard.” Diperbarui 30 September 2026.

142. Wang, J., A. Hoecker, A. Koh, dkk. *Managing Token Spending without Choking Off Opportunity*. Bain Technology Report 2026, 29 September 2026. https://bain.com/insights/managing-token-spending-without-choking-off-opportunity-technology-report-2026/.

143. Kwa, T., B. West, J. Becker, A. Deng, K. Garcia, dkk. “Measuring AI Ability to Complete Long Software Tasks.” METR. arXiv:2503.14499v3, 25 Februari 2026.

144. Doerer, Kristen. “Klarna credits AI for slashing customer service costs.” *CX Dive*, 21 Mei 2025. https://www.customerexperiencedive.com/news/klarna-ai-slash-customer-service-costs/748647/.

145. Doerer, Kristen. “Klarna Reinvests in Human Talent for Customer Service.” CX Dive, 9 Mei 2025.

146. Wang, Yiwei, Chuan Zhu, Tianjun Feng, dkk. “Agentic AI and Human-in-the-Loop Interventions: Field Experimental Evidence from Alibaba’s Customer Service Operations.” Pracetak, arXiv 2605.14830, versi 1 terbit 14 Mei 2026, versi 2 terbit 1 Juni 2026. https://arxiv.org/abs/2605.14830.

147. Gartner. “Hype Cycle for Agentic AI.” 15 April 2026. https://gartner.com/en/articles/hype-cycle-for-agentic-ai.

148. Cemri, Mert, Melissa Z. Pan, Shuyi Yang, Lakshya A. Agrawal, dkk. “Why Do Multi-Agent LLM Systems Fail?” arXiv:2503.13657v3, 26 Oktober 2025.

149. Doddapaneni, Purna, Chris McLaughlin, Stanley Liu, dan Bill Radzevych. “The Missing Architecture for Agentic Software Development.” Bain Technology Report 2026. Bain & Company, 29 September 2026.

150. “BRI Gandeng Microsoft Tingkatkan Pemanfaatan AI untuk Inklusi Keuangan.” Investortrust, 27 April 2024. https://investortrust.id.