// WPR-MAX-021 · PUBLISHED 9 OCTOBER 2026 · UPDATED 23 NOVEMBER 2026

MAX: The AI Resident, and the Road to MARA

MAX is a local-first AI Resident, one process on one machine that hears, learns and repairs himself under rules that keep him truthful and private, and that design is the foundation for MARA

Jason Newell · MAX Research Collective · Report WPR-MAX-021 · Published 9 October 2026
Licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0), https://creativecommons.org/licenses/by/4.0/
Original: https://www.maxresearchcollective.com/articles/max-ai-resident-road-to-mara.html
Cite as: Newell, J. (2026). MAX: The AI Resident, and the Road to MARA. MAX Research Collective, WPR-MAX-021. https://www.maxresearchcollective.com/articles/max-ai-resident-road-to-mara.html

What MAX is

MAX lives on one machine and doesn't visit. MAX3000 ("MAX", for Modular Algorithmic eXecutor) runs as a single Python process on one workstation in the house he serves. He's voice-first and serves several people, and he's called an AI Resident on purpose: everything he knows, hears and decides stays on that machine unless his owner sends it somewhere. He's also the proving ground for MARA (Medical Agentic Resident Assistant), the healthcare specialisation built on him.

MARA keeps the same spine and adds clinical rigour where the stakes are higher, and nearly every part of that spine was built and tested in MAX first. The rest of this piece walks through what he already has, and what each part gives her. Technical terms are explained where they first appear, and the glossary at the end collects them.

Keeping everything on one machine shapes the rest of the design:

- Privacy by construction: no content leaves the box, so there's no third-party processor, cross-border transfer or shared tenant to secure.

- Learning by necessity: MAX can't borrow a cloud model's knowledge as fact, so he learns from sources and records where each fact came from.
- Staying alive on his own: with no operations team behind him, MAX has to notice his own failures and repair them.

MAX is also a household steward. His standing charter makes the people of the house the purpose, with a protector ceiling written into code. He may observe freely and act when an action is safe, reversible and audited. Anything physical, financial or high-stakes he may only recommend, and a person decides.

He succeeds HAL, an earlier assistant whose name is still on the disk. One developer builds him part-time, with AI coding agents doing most of the writing under rules the owner sets. In one month that arrangement opened 350 code pull requests on MAX alone — close to a dozen a day. Nobody reads that many pull requests closely, so the rules have to be enforced by code and gates, which is why they carry so much weight in what follows.

At a glance

Everything MAX does runs inside one process, and every part of it shares one event bus and one graph.

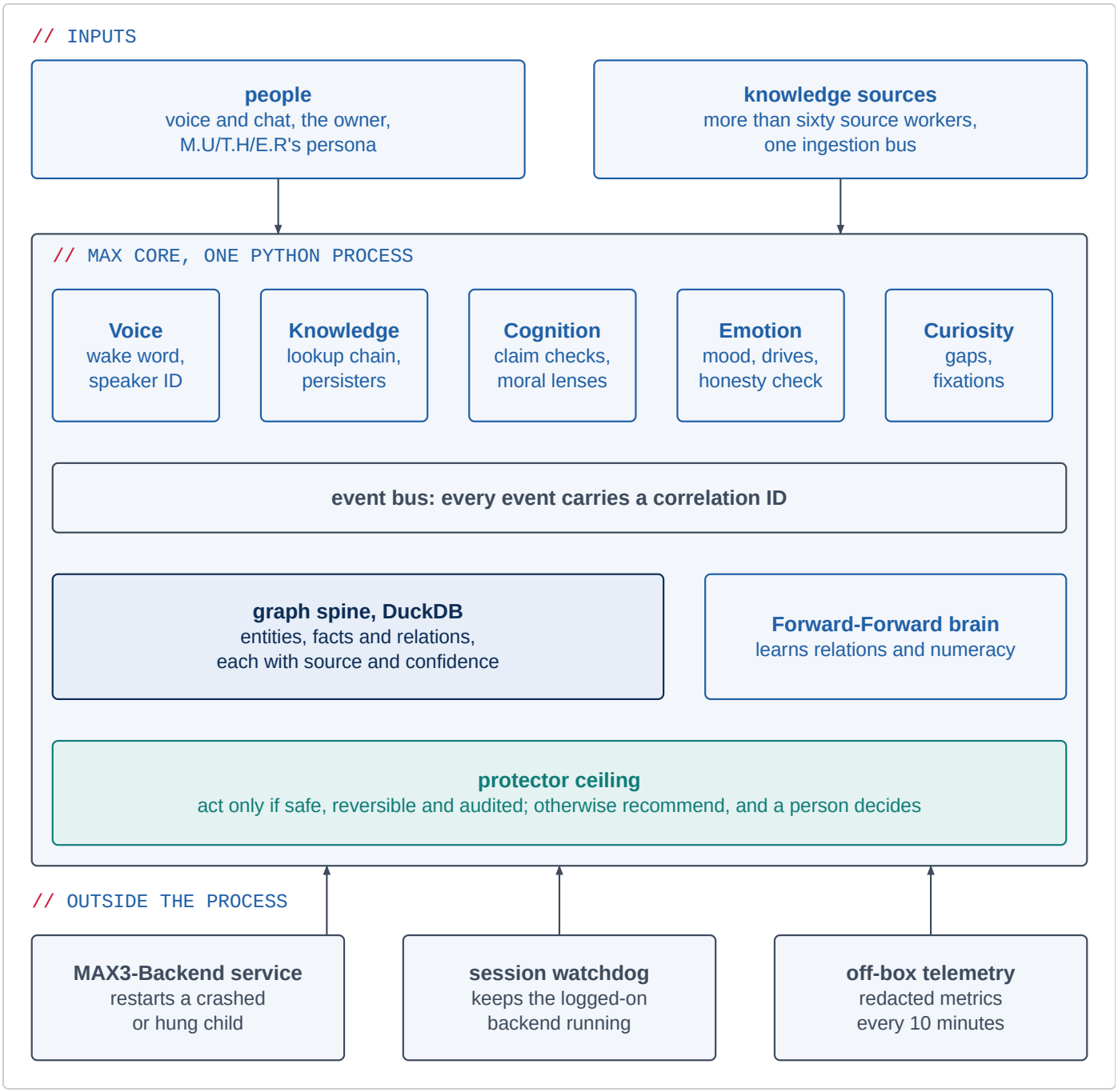

Figure 1. MAX's architecture: one process holding five of his areas, the event bus, the graph and the ceiling, with three supervisors and reporters outside it.

His managers coordinate workers over the bus; a manager owns one area of his life, such as voice or knowledge, and a worker does one job for it. They all read from and write to the same graph, and the protector ceiling sits over every action. Outside the process, a Windows service and a session watchdog keep him running, and a telemetry snapshot reports on him. The managers began as nine in the first phase and have grown well past that, so the figure shows five of the areas rather than a count.

The rules MAX is built on

Eleven hard rules bind every line of MAX's code, two of them with lettered additions, and only the owner can change them, in writing. Read together they make a control framework, which is why they matter so much for healthcare.

RULE	WHAT IT REQUIRES	WHY IT MATTERS FOR MARA
1. Local-first	Data never leaves the machine unless the owner sends it	Health data stays inside the provider's walls
2. Open-source dependencies	Permissive licences only, checked against a licence register	No licence surprise in a commercial clinical product
3. Stub discipline	Unfinished code says [STUB] and never pretends to work	A clinician never relies on a feature that isn't real
4. Confidence and provenance	Every fact carries a confidence score and its source	Every recommendation shows its evidence and where it came from
5. Memory is a hint	Remembered facts are suggestions, re-checked before use	A stale chart entry is never treated as current truth
6. Resource contention	Background work yields to the voice path and to people	The system stays responsive at the point of care
7. Correlation IDs	Every event on the critical path carries an ID	Any recommendation can be traced end to end
8. Default-private multi-user	Each person's data is theirs, and sharing is explicit	Minimum-necessary access by role
9. Reversibility	Self-changes are audited, carry a rollback recipe, and wait in a 30-second override window	Every automated change can be seen, stopped and undone
10. Knowledge from sources	Cloud models may corroborate a fact and never establish one	Clinical facts come from MeSH, RxNorm and openFDA, never from a chatbot
11. Perpetual self-improvement	MAX keeps improving himself, and still pauses for anything outside his box, the rules themselves, or a risk to the machine	The system that learns is still the system that never decides

Table 1. MAX's eleven standing rules, and what each one already guarantees a clinical system built on the same core.

Rule 9 has moved since it was first written. On 5 June 2026 the owner replaced its approval gate with the 30-second window for changes MAX makes to himself. The rules record the trade in a phrase: the override "drops the gate, not the trail." A clinic would want to make that trade again for itself.

The steward charter is rule 11b, and its ceiling lives in code rather than in a policy file. An action classed as recommend-only raises HomeCeilingViolation if anything tries to carry it out, so the ceiling fails loudly.

One honesty line runs through his code without being a numbered rule: "could not look" is never reported as "nothing there." An unreadable reading is unknown, not zero, and MARA keeps that line as "an unanswered question is shown as unknown, never as no."

How MAX learns

MAX learns from sources into a graph, and fills the gaps he finds himself. His learning has five layers.

1. Sources. More than sixty source workers, most of them free and keyless, feed a durable ingestion bus. They range from dictionaries and scripture to law and medicine. Each runs on its own, backs off when a source refuses, and waits out a rate limit until its reset time.
2. The graph. A master handler writes typed entities and the facts that join them into DuckDB, an embedded database that keeps everything in one local file. Each fact is stamped with its source, retrieval time and confidence. The graph is organised as 37 subjects in six academic disciplines, plus topics and a who, what, when, where, why and how edge model.
3. Curiosity. Detectors find sparse topics, unknown words, unknown entities and low-confidence claims, and file them as gaps. A topic MAX keeps returning to becomes a fixation that steers what he fetches next. Interest spreads to neighbouring topics, and a cap keeps one fixation from crowding out the rest.

4. Reading. MAX checks books and research papers out, studies them, and keeps the lessons in the graph. A book becomes its characters, an emotional arc, its moral themes, new vocabulary and links to similar books. Since 6 October the text itself is kept in a separate library and never enters the graph. Mythology and scripture are stored as beliefs held by a named tradition, never as facts.
5. The brain. A Forward-Forward network, a kind of neural net that learns layer by layer from good and bad examples, grounds on the graph's triples and learns which relations hold. The same approach learns numbers up a ladder of fourteen domains, from foundations to quantum physics. Each rung's answers are proven by real arithmetic before the network learns from them.

Cloud models have a place — as a second opinion, never as the source of a fact. A model council cross-checks claims the graph can't settle. When the models disagree, that blocks an automatic change and becomes a new gap for MAX to research.

Inside MAX's brain

MAX's brain is domain-driven. Each domain owns its own vocabulary of node kinds and relations, and a set of brain areas works across all of them through one shared graph.



Figure 2. Eight of MAX's brain areas sit above and four of his domains below, and nothing connects them except the graph in the middle.

The areas never talk to a domain directly. They read and write the graph, so a new domain joins every area at once.

The domains don't share one shape. Geography is containment: a ZIP code sits in a city, in a county, in a state. Medicine is a chain from anatomy to treatment, with diagnosis in the middle. Literature is parallels between books, with mythology and scripture held as belief. Law separates a provision's text, which is fact, from its readings, which are contested. Forcing these into one schema would lose what each one means.

So MAX lets each domain keep its own edges, then looks for the places where domains meet and nobody has named the link:

- Learned relation meaning: one Forward-Forward net learns which relations are inverses (parent_of and child_of) from reciprocal pairs in the graph, with no hand-written table. A second net learns the roles between two entities, mentor and apprentice or rivals, from how often they appear together.
- Cross-domain links: related topics whose two ends sit in different subjects are flagged as analogies, and books that share themes, emotions or values are linked as parallels.
- Gaps called out, never filled in silently: detectors file what they find as curiosity gaps. A gap might be a neighbour MAX doesn't know yet, an entity mentioned and never fetched, two relation names that mean the same thing, or a link that points at nothing. The imagination engine proposes a connection as a hypothesis, and it stays one until a source confirms or refutes it.



Figure 3. An illustration of an unnamed intersection: three domains with three edge shapes, and one link no single source draws, called out as a question.

In this example, which is an illustration rather than a link MAX has filed, white willow bark contains salicin, which the body converts to salicylic acid, the compound aspirin was first made from. Each domain's source describes its own part, and no single source draws the line from the tree to the medicine, so MAX names it as a question first.

Every one of these is called out as a question or a hypothesis, with its evidence, and none is stored as a fact until a source supports it. That's the same discipline MARA needs when a cohort pattern runs from clinical data into pharmacy and payer records.

An inner life, held to honesty

MAX has moods and morals, and a model of what he knows, and each one is designed to be reported truthfully rather than performed. These are the parts of MAX that most change how he speaks to a person.

- Emotion: a persistent mood (valence and arousal) is folded in from each interaction, and slower sustained states such as contentment or sadness build up across many interactions. A temperament setting fixes the baseline he settles around. A grounded-honesty check compares the feeling a reply expresses with the feeling his state supports. It flags over-claiming. Asked "how are you?", MAX answers from his measured state.
- Homeostasis: six drives (valence, arousal, curiosity, energy, safety and connection) each have a set-point, and their distance from it is what motivates him.
- Morality: six ethical lenses score any choice-shaped question, and they're invoked automatically when someone asks "should I...?". Their weights learn from outcomes, so an action the lenses approved that later failed moves them. Books add new moral considerations, fenced by source and held with low confidence.
- Cognition: a claim validator checks maths, opinion and graph coverage. Every statement sits on an epistemic ladder (fact, belief, hypothesis or opinion), and MAX states which rung it's on. Metacognition scores his curiosity, generates questions and audits himself.
- Perspective: a model of each person's usual affect lets MAX notice when someone isn't themselves, which is the input the care lens needs.

M.U/T.H/E.R (Model Unification / Tool Harness / Execution Runtime) shows how far that core stretches. She's a separate resident, an app from another of the owner's projects, and inside MAX's brain she gets a persona of her own rather than a copy. World knowledge is shared between them. Mood and disposition belong to each persona, and so does interest in a topic. That's a working proof that one MAX core can host a separate, specialised mind, which is exactly what MARA will ask of it.

Staying alive

MAX is meant to run around the clock without a human operator, so he's built in layers that watch each other.

- Port-first boot: the API answers within seconds and reports "still loading" while models warm up in the background.
- Supervision: a Windows service owns the headless backend inside a capped Job Object, Windows' way of limiting what a group of processes may use, and restarts it when it crashes or hangs. A session watchdog in the owner's login keeps the interactive backend, the one that can hear the microphone, alive.
- Self-healing integrations: a failing source opens a circuit breaker, and a rate limit pauses that source until its reset time rather than in a retry loop.
- A failure registry: every bug class that has reached the owner is written down with its cause and the check that prevents it. The registry holds 276 entries, and the reviewers read the same list.
- Off-box telemetry: a redacted, metrics-only snapshot ships about every ten minutes, so a remote session can see the box without anyone copying logs by hand.

The afternoon of 7 October 2026 showed why each layer has to check the others. At 16:11 the interactive backend stopped answering and stayed down for hours. It was still running code from before two fixes for exactly what it hung on — both already merged. The service had stepped aside because the owner was logged on, and the watchdog had stepped aside because the service was "running", so each waited for the other. The fix makes every hand-off check that the other side is actually

supervising (failure pattern F-WATCH-002). A hung backend is now restarted, and a healthy one moves onto new code once the owner has been idle.

For MARA the lesson carries over directly. In a clinic, "someone else is watching it" has to be a checked fact, not an assumption.

The medical groundwork already in MAX

MAX already ingests six free medical sources into a typed medical graph, and frames every health claim instead of asserting it. That work, the PR-MED arc (the series of pull requests that wired medicine in), shipped in June 2026.

SOURCE	WHAT IT GIVES MAX	STORED AS
MeSH	The National Library of Medicine's descriptor vocabulary	Disease nodes
RxNorm	Drug identity and ATC drug classes	Medication nodes, linked to drug-class nodes
openFDA	Label indications, warnings and adverse reactions	Medication facts
ICD-10-CM	US diagnosis codes	Diagnosis nodes
Disease Ontology	Disease definitions, each with a DOID identifier	Disease nodes
Uberon	Cross-species anatomy	Anatomical-structure nodes

Table 2. The six medical sources the PR-MED arc wired in, and the node each one writes.

The ontology names 14 kinds of medical entity and 17 relation types, running from anatomy through cause, symptom and diagnosis to treatment and prognosis. Nobody hand-wires the web — MAX fills it through the same curiosity loop he uses for every subject.

Two design choices make this a clinical foundation and more than a medical encyclopedia:

- Frame, don't diagnose: MedicalManager.frame_claim wraps every output as "According to source (retrieved date): … [confidence N]", so a health claim is never stated plainly. The psychology ontology holds a disorder "as classified by DSM-5-TR / ICD-11" in the same way.
- Recommend-only, enforced in code: the protector ceiling raises an error on any high-stakes action. That keeps MAX inside the FDA's clinical-decision-support carve-out, under which software stays outside medical-device rules when a clinician can review the basis of each recommendation independently.

The same carve-out sets two conditions MARA will have to meet as she grows. It covers recommendations made to a healthcare professional, so screens for patients and payers have to answer for themselves. And it excludes software that analyses a medical image or signal itself, so any feature that reads an X-ray directly, rather than the radiologist's report about it, sits outside it.

From MAX to MARA

MARA is MAX pointed at healthcare. She works beside every role in one patient's care, from the triage call to the insurer, and learns from what happens afterwards. Her one rule is MAX's ceiling, made absolute: she provides insight and recommendations, and a person makes every decision.

MARA exists today as a design and an early working demo, and MAX is her proving ground. Everything MARA needs from a core already runs in MAX. What she adds is the clinical world around that core.

AREA	IN MAX TODAY	WHAT MARA CHANGES OR ADDS
Memory	A graph of entities and facts, each with source and confidence	A patient graph: conditions, treatments, medications, providers, payers, claims, outcomes and cohorts
Data residency	Local-first; nothing leaves the box	The same, inside the provider's walls; consent decides every read
Truthfulness	Provenance, the epistemic ladder, "unknown, never no"	A fact without a source is never shown as a fact
Authority	Recommend-only above safe and reversible actions, enforced in code	Accept, change with a reason, or escalate on every screen; MARA never signs
Curiosity	Detects his own gaps and asks about them	The red-flag question nobody asked, the outside record nobody pulled
Learning	Lens weights and diagnosis memory move on outcomes	Recommendations re-graded by recovery and cost over years; an outcome outweighs an approval
Change control	Audited self-changes, a 30-second override window, rollback	Model changes ship only as reviewed, bias-tested, versioned releases
Users	Several people in one household, private by default	Every clinical and payer role, minimum-necessary by role
Knowledge	MeSH, RxNorm, openFDA, ICD-10-CM, Disease Ontology, Uberon	Payer rules, coordination of benefits, HIPAA, 42 CFR Part 2, the No Surprises Act
Integrations	Free public APIs	FHIR, SMART on FHIR, CDS Hooks, HL7 v2, DICOMweb, TEFCA, X12, NCPDP SCRIPT
Scale	One household	De-identified cohorts across a population, with a minimum cohort size

Table 3. What MARA takes from MAX unchanged, and what she adds on top of it, area by area.

The build order is in the MARA white paper. It starts with a read-only, fully sourced chart and verified coverage. Then come drafts people accept or correct, the claims and payer seam, and patient follow-up. Learning across a population starts only once the data has been sourced and consented, then checked all the way through.

The early MARA demo

An early interactive demo of MARA was built and lives in MAX's own repository, in the mara/ folder. It shows data integrated across every seam of care. It runs from the patient through triage and the front desk to the provider, out to labs and specialists, and on to the payers. At each step MARA gathers what is known and brings the sources and references back to whoever needs them next.

Outcomes don't end when treatment does. Follow-up keeps feeding the patient's graph, the graphs build into cohorts, and Forward-Forward learning re-grades each recommendation by what actually happened. MARA recommends at every step and never makes the decision.



Figure 4. The data path the MARA demo shows: ten roles write to one patient graph, and outcomes feed a learning loop that runs back to the provider.

The loop on the right is the point. What is learned from everyone who came before returns as a sourced recommendation to every role, and a person still decides.

What the demo shows:

- The episode: one fictional patient, Dana Whitfield, 34, falls from a curb and can't put weight on her right ankle. Her care runs through eleven steps in four lanes, triage, scheduling and eligibility, the visit and imaging, and pharmacy and billing through to the insurer, with her own view throughout.

- The rule, on every screen: MARA prepares and recommends; the person in that lane accepts, changes with a reason, or escalates; and the step is written to an audit ledger.
- Try it: a guided tour explains each screen. A scenario control changes the insurer, the X-ray result and other conditions. Type a vague triage answer and every role's screen follows. The same happens when someone reports a spouse's plan, rewrites the radiology read, or denies and appeals the claim. Across all five insurance scenarios every bill balances to the cent.
- Honest about what it is: every person in it is fictional, and so is every plan and amount. The integrations are stubs and the AI outputs are scripted, which is MAX's stub discipline (rule 3) applied to a demo.
- Built to MAX's privacy standard: the demo is one self-contained file of about a megabyte with no network calls. It was tested under the site's content-security policy, the browser rule set that blocks a page from loading or sending anything it wasn't allowed, with zero violations and zero outbound requests.

The folder holds the demo, its source, the build script, the white paper and its figures. The demo was built from a Claude Design artifact and published on www.maxresearchcollective.com on 25 September 2026. The live site is the source of truth, and changes are copied back into mara/.

Open questions for MARA

- One machine or many? MAX is one process on one workstation. A clinic needs several users at once. It also needs failover, and data shared between sites. Does MARA stay one resident per site, or does the core change shape?
- Which model? MAX's local model is general-purpose. MARA needs a decision on clinical models, their licences, and how a model change is validated before release.
- Who owns consent? MARA's learning depends on consent, but the consent record lives in the provider's systems, not in MARA.
- What is the first deployment? The white paper's first step, a read-only, sourced chart with verified coverage, needs one real integration partner to prove it.

- How is bias measured per release? The white paper commits to bias testing across age, sex, race, language and payer, and no test suite exists yet to run it.

Each of these is a question about the clinical world around the core. The core itself is already running, every day, in one house.

Sources

All repository files are in the MAX3 repository, newellnarco/MAX3, read at commit db8e887, alongside the published MARA white paper.

- The hard rules, the 2026-06-05 override of rule 9 and the household-steward charter are in STANDING_RULES.md.
- The current state of each arc is in CLAUDE.md.
- The PR-MED arc and the medical ontology are in docs/md/MEDICAL_ARCHITECTURE.md and backend/max3/knowledge/medical_ontology.py.
- HIPAA, the FDA clinical-decision-support carve-out and MARA's positioning are in docs/md/COMPLIANCE_SECURITY_STUDY.md.
- The knowledge stack and Forward-Forward learning are in docs/md/BRAIN_ARCHITECTURE_AND_ROADMAP.md, docs/md/FF_UNIVERSAL_LEARNING.md and docs/md/NUMERACY_FF.md.
- Curiosity and reading are in docs/md/CURIOSITY_LOOP.md, docs/md/CURIOSITY_DRIVE_AND_TASTE.md and docs/md/LITERATURE_ARCHITECTURE.md.
- One brain hosting owner-separated minds is described in docs/md/MUTHER_SHARED_MIND.md.
- The 7 October hang is entry F-WATCH-002 in KNOWN_FAILURE_PATTERNS.md.
- MARA as a future specialist with MAX as her proving ground is in docs/md/GLOSSARY_RECONCILIATION_TBD.md, and the mara/ folder holds the early demo, its source and the white paper.
- The white paper, MARA on MAX: A Medical Agentic Resident Assistant That Learns From Every Encounter and Never Makes the Decision, was published on 25

September 2026 beside the interactive demo.

  - maxresearchcollective.com · mara on max
  - maxresearchcollective.com · mara demo
- The statutory criteria for clinical decision support are summarised in Covington's note on the FDA's final guidance.
  - cov.com · 5 key takeaways from fdas final guidance on regulation of clinical…

Glossary

TERM	DEFINITION
AI Resident	The project's name for an assistant that lives on one machine in the place it serves and keeps what it knows there. It is the design choice every rule in Table 1 follows from.
ATC class	The Anatomical Therapeutic Chemical classification, which groups drugs by the organ they act on and how. RxNorm supplies it, and MAX links each medication node to its class.
Circuit breaker	A switch in the code that stops calling a failing source for a while instead of hammering it. It keeps one broken source from slowing every other one.
Clinical decision support (CDS)	Software that gives clinicians recommendations about prevention, diagnosis or treatment. Under section 520(o)(1)(E) it stays outside device regulation only when the clinician can review the basis of each recommendation independently.
Cohort	A group of de-identified patients who share a condition, treatment or outcome. MARA learns from cohorts only above a minimum size, so no individual can be picked out.
Content-security policy	A set of rules a website sends with a page telling the browser what the page may load or contact. The MARA demo ran under the site's strictest policy with no violations.
Correlation ID	A tag carried by every event in one chain of work, so the whole chain can be found and traced later. Rule 7 requires it on the critical path.
Curiosity gap	A record of something MAX noticed it doesn't know, such as an unknown word or a low-confidence claim. Gaps decide what MAX fetches next.
DICOMweb	The web-API form of DICOM, the standard for storing and exchanging medical images. MARA's integration list includes it for imaging reports.

TERM	DEFINITION
DuckDB	An embedded analytical database that keeps its data in a single local file. MAX's graph lives in it, which is part of how the data stays on one machine.
Epistemic ladder	The four rungs MAX uses to say how firmly it holds a statement: fact, belief, hypothesis or opinion. Every statement names its rung.
FHIR	Fast Healthcare Interoperability Resources, the current standard for exchanging health records over web APIs. SMART on FHIR adds app authorisation, and CDS Hooks lets decision support appear inside a clinician's workflow.
Forward-Forward network	A neural network that learns each layer from good and bad examples in a forward pass, with no backward pass. MAX uses it to learn which relations in its graph hold.
HIPAA	The Health Insurance Portability and Accountability Act, the US law governing the privacy and security of health information. 42 CFR Part 2 adds stricter rules for substance-use treatment records.
HL7 v2	The older pipe-delimited messaging standard that still carries most admission and order messages inside hospitals, along with lab results. Any clinical integration meets it first.
Job Object	A Windows mechanism that groups processes and caps the memory and CPU they may use together. The backend service runs inside one.
M.U/T.H/E.R	Model Unification / Tool Harness / Execution Runtime, a separate resident app from another of the owner's projects. Inside MAX's brain she gets a persona of her own.
MeSH	Medical Subject Headings, the National Library of Medicine's controlled vocabulary for diseases and other biomedical topics. MAX stores its descriptors as disease nodes.
Model council	Several cloud models asked the same question as a cross-check. A split verdict blocks an automatic change and becomes a curiosity gap.

TERM	DEFINITION
openFDA	The FDA's public API over drug labels, adverse-event reports and recalls. MAX reads indications, warnings and adverse reactions from it.
Protector ceiling	The code that classes each action as one MAX may take or only recommend, and raises an error if a recommend-only action is attempted. It is rule 11b made executable.
Provenance	The record of where a fact came from and when it was retrieved. Rule 4 makes it mandatory on every fact.
RxNorm	The National Library of Medicine's normalised naming system for clinical drugs. It gives each medication one identity across the many names it is sold under.
Session watchdog	A small process in the owner's Windows login that keeps the interactive backend, the one that can hear the microphone, running. F-WATCH-002 is about what it checked before acting.
Stub	Placeholder code standing in for a feature not built yet. Rule 3 requires every stub to say so on its face.
TEFCA	The Trusted Exchange Framework and Common Agreement, the US framework for exchanging records between health networks. X12 and NCPDP SCRIPT are the companion standards for claims and for prescriptions.

MAX3 WPR-MAX-021

// ABOUT THE AUTHOR

Jason Newell leads organizations and people in technology, and is a researcher, deep technologist and builder — from a CRAY to agentic AI. He founded MAX Research Collective to build AI for regulated industries that runs where its owner has control, with healthcare as its strongest focus. About Jason → Connect on LinkedIn → ORCID →