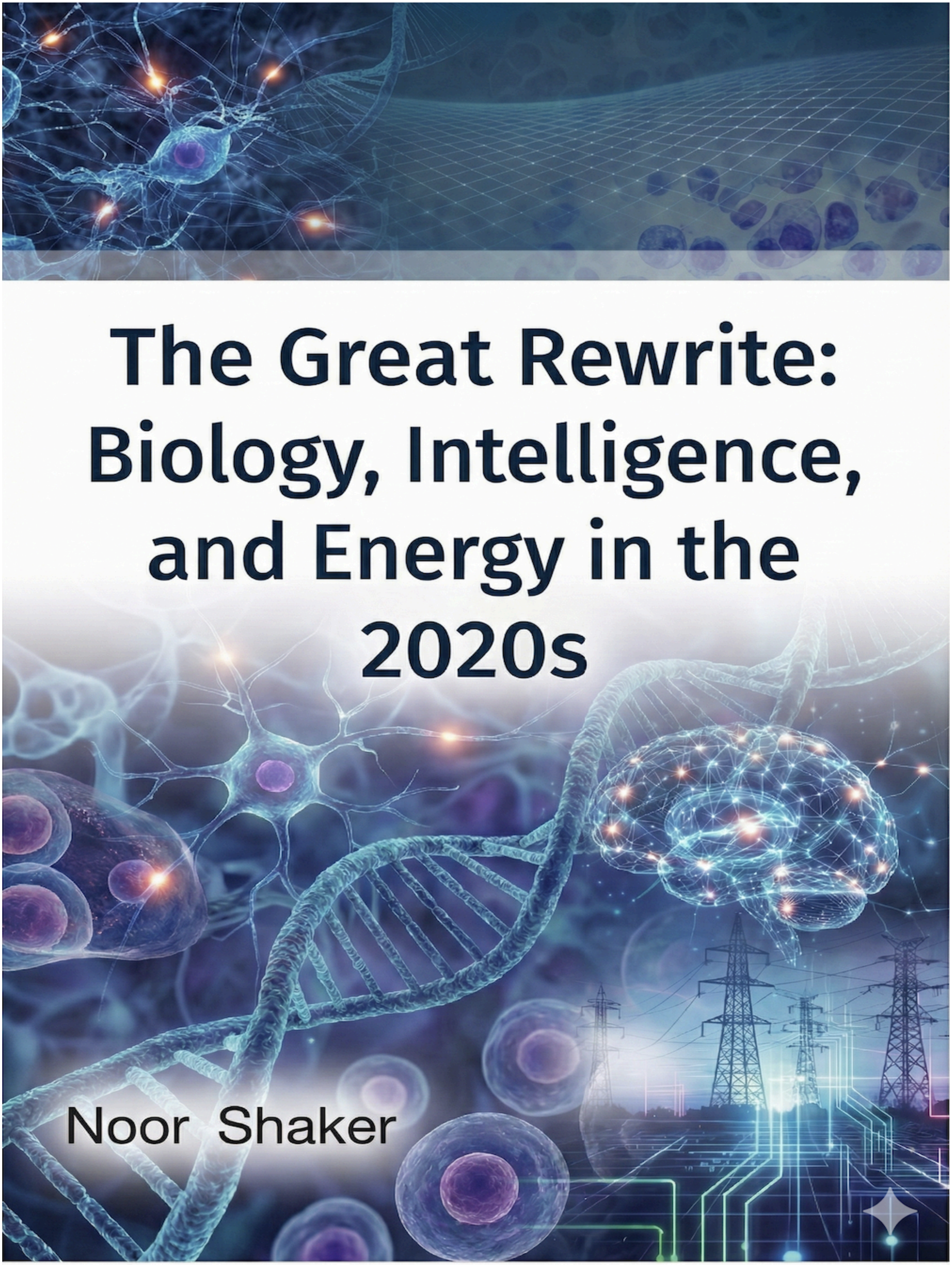

The Great Rewrite: Biology, Intelligence, and Energy in the 2020s

Noor Shaker

The Great Rewrite

Biology, Intelligence, and Energy in the 2020s

Noor Shaker

بسم الله الرحمن الرحيم

{اقْرَأ}

Noor Shaker

Press

Preface

For most of history, human capability advanced through the accumulation of knowledge and the refinement of tools. We learned to understand the world through observation. We developed technologies that amplified human labor, extended human senses, and accelerated human communication. But the fundamental constraints remained: biology limited what our bodies could do, intelligence required human cognition to direct every decision, and energy systems operated within the bounds of what previous generations had built.

Between 2020 and 2025, something changed. Decades of fundamental research across multiple scientific domains—molecular biology, machine learning, materials science, renewable energy engineering—reached maturity simultaneously. Manufacturing processes that had been confined to laboratories scaled to commercial production. Technologies that had been limited to controlled demonstrations achieved sufficient reliability for real-world deployment. And capabilities that had seemed perpetually five to ten years away crossed thresholds into operational reality.

The result is a transformation that extends across the biological, the cognitive, and the physical infrastructure of civilization itself.

In medicine, we moved from managing disease to engineering cellular function. In artificial intelligence, machines transitioned from answering questions to autonomously executing complex tasks. In energy, renewable sources surpassed fossil fuels not through policy mandates but through economic superiority. Each shift represents decades of incremental progress suddenly converging into deployed capability.

This book documents convergence through ten detailed narratives spanning 2020 to 2025. The structure reflects three domains undergoing simultaneous transformation:

Part I: The Rewrite of Biology follows medicine's transition from reactive management to molecular engineering—preventing HIV transmission entirely rather than suppressing it with daily medications, generating replacement organs from stem cells rather than waiting for donors, bypassing severed neural pathways with brain-computer interfaces rather than accepting permanent paralysis.

Part II: The Intelligence Shift documents artificial intelligence crossing from controlled environments into autonomous operation—vehicles navigating city traffic without human oversight, AI systems planning and executing multi-step tasks independently, and models compressed from data centers to smartphones while maintaining capability.

Part III: The Planetary Substrate traces transformations in energy and observation—renewable electricity surpassing coal globally, aviation beginning decarbonization through sustainable fuels, and astronomy shifting from taking occasional snapshots to filming the universe continuously.

These stories share common characteristics. Each represents an inflection point of learning curves that spanned decades. Each required manufacturing at scales that seemed impossible until accomplished. Each crossed thresholds where capabilities once limited to demonstrations became commercially deployed or scientifically operational. And each demonstrates that constraints once considered fundamental—biological limits, cognitive boundaries, physical infrastructure—can be renegotiated when underlying technologies mature and supporting systems are built.

The transformations documented here emerged from thousands of researchers making incremental advances, engineers solving manufacturing challenges, entrepreneurs building companies around maturing technologies, and policymakers creating frameworks that enabled deployment.

The ten chapters that follow document the technologies I identified as most fundamental, most transformative, and most clearly at inflection points between 2020 and 2025. These are not predictions about what might happen but records of what has already occurred—the capabilities now deployed, the thresholds already crossed, the systems currently operational. Understanding what happened in these five years is essential to understanding what becomes possible in the decade ahead.

Enjoy reading,

Noor

Part II: The Intelligence Shift

Noor Shaker

AI & Automation

For most of computing history, software has been an exercise in explicit instruction. Programmers identified problems, decomposed them into logical steps, and wrote code that executed those steps precisely. When programs encountered situations their creators hadn't anticipated, they failed. When tasks required understanding context, recognizing patterns, or adapting to variation, they required human intervention at every decision point.

Between 2020 and 2025, artificial intelligence crossed multiple thresholds simultaneously. Systems that had been limited to answering questions began planning and executing multi-step tasks autonomously. Models that required data center infrastructure were compressed to run on smartphones without sacrificing capability. Search engines that returned lists of links began generating direct answers and executing complete solutions. And machines that had been helpless in the physical world—unable to navigate unpredictable environments or manipulate irregular objects—achieved sufficient competence to operate commercially in urban traffic and industrial facilities.

The two chapters that follow document this transformation across the digital and physical domains. Each represents a distinct technical challenge. Each emerged from decades of foundational research in machine learning, computer vision, and robotic control. But together, they reveal a unified shift in what machines can accomplish without human guidance.

The two chapters of Part II document this transformation through specific technologies and deployment milestones. Chapter 6 follows the path from controlled laboratory demonstrations to commercial autonomous vehicles transporting hundreds of thousands of passengers daily, and from rigid industrial robots to fast-learning systems that adapt to new tasks through observation and simulated practice. Chapter 7 traces the architecture shift from conversational AI to agentic systems, the compression of models from data centers to devices, and the death of link-based search in favor of generated answers and executed solutions.

Chapter 6: The Physical Mind

Robotaxis and Fast-Learning Robots

Noor Shaker

The 100,000th Ride

On March 15, 2025, Jennifer Valdez settled into the driver seat of a white Jaguar I-PACE as rain drummed against the windshield. She was not really driving. The steering wheel turned on its own, guided by eight cameras, five lidars, and six radars processing 1.4 million data points per second. This was her Waymo robotaxi—and this particular ride, departing from San Francisco International Airport at 6:47 PM, was not her first trip with the service.

Valdez, a pharmaceutical sales representative, had been skeptical when Waymo opened its San Francisco service to the public in 2024. She'd spent years driving herself to client meetings, convinced that no machine could navigate the city's chaotic intersections, aggressive drivers, and impossible parking. Her first robotaxi ride was curiosity mixed with terror.

By the end of 2025, Waymo was providing 450,000 autonomous rides weekly across San Francisco, Los Angeles, Phoenix, and Austin—roughly 64,000 rides per day, or one every 1.4 seconds. The company had logged over 50 million fully autonomous miles. The accident rate, measured per million miles, was 85% lower than human drivers.

But the real story wasn't just about removing human drivers from cars. It was about machines finally learning to navigate the physical world with something approaching human intuition—and learning to do it faster and safer than anyone thought possible.

October 9, 2010: The DARPA Challenge

To understand how Jennifer Valdez came to trust her life to a driverless car, we need to rewind to a moment when the entire concept seemed like science fiction hovering just beyond reach.

In November 2007, six vehicles successfully completed the DARPA Urban Challenge, navigating ~60 miles of simulated city traffic near Victorville, California. Carnegie Mellon’s Tartan Racing team won with 'Boss' (a modified

Chevy Tahoe), while Sebastian Thrun’s Stanford team took second with 'Junior' (a Volkswagen Passat).

The victory was impressive—but also revealed the technology's profound limitations. Boss required a trunk full of computers, sensors, and custom electronics drawing significant power (~1–2 kW total system load). The course had been carefully controlled, with speeds limited to 30 mph and all other vehicles operated by professional drivers following strict DARPA protocols. There were no pedestrians, no bicycles, no unexpected obstacles. The weather was clear and dry. This wasn't real-world driving—it was a highly constrained demonstration.

Thrun, a German-born computer scientist who had joined Stanford's faculty in 2003, understood both the achievement and the chasm that remained. Three years after the DARPA Challenge, in 2010, Thrun joined Google as a founding director of Google X, the company's secretive research division focused on "moonshot" projects. His first major initiative was Project Chauffeur—what would eventually become Waymo.

The project started with a modified Toyota Prius equipped with a spinning Velodyne lidar unit on the roof—the distinctive cylinder that would become the visual signature of early self-driving cars. The hardware was crude by today's standards: 64 laser beams rotating at 10 Hz, creating a 360-degree point cloud of the vehicle's surroundings. Each lidar unit cost approximately $75,000.

But the real challenge wasn't hardware. It was software—specifically, teaching a machine to make sense of the chaos of real-world driving.

The Perception Problem

Driving seems simple because human brains make it look effortless. You glance at a busy intersection and instantly parse a staggering amount of information: the traffic light is yellow, the pedestrian at the crosswalk is looking at their phone and probably won't cross, the delivery truck on the right is partially blocking the bike lane, the cyclist behind the truck is about to swerve left, the sedan three cars ahead just hit its brakes.

Your brain does this through pattern recognition honed over millions of years of evolution and refined through hours of driving experience. You don't consciously process "object at 15 meters, trajectory intersecting path, classification: probable hazard." You just see and react.

For a machine, every element of this scene is a computational nightmare.

Early self-driving systems used rule-based approaches. Engineers would write explicit code: "If object appears in path and closing speed exceeds X, apply brakes with force Y." These systems worked in controlled environments but failed catastrophically when encountering situations the programmers hadn't anticipated.

The breakthrough came from a different approach entirely—not programming rules, but learning from examples.

The Neural Network Revolution

The conceptual foundation for modern AI traces back to 1943, when Warren McCulloch and Walter Pitts published "A Logical Calculus of the Ideas Immanant in Nervous Activity," proposing that networks of simple computational units could perform complex logical operations—essentially, that artificial neurons could think.

But the real revolution began much later, in 2012, when Geoffrey Hinton's research group at the University of Toronto stunned the computer vision community at the ImageNet Large Scale Visual Recognition Challenge.

ImageNet was a dataset containing 1.2 million labeled images across 1,000 categories—everything from golden retrievers to garbage trucks to broccoli. The competition challenged teams to build systems that could correctly classify these images. Prior to 2012, the best systems achieved error rates around 25%—meaning they misidentified one in four images.

Hinton's team, using a deep convolutional neural network called AlexNet, achieved an error rate of 15.3%. It wasn't an incremental improvement—it was a different category of performance. Within two years, deep learning systems would surpass human-level accuracy on ImageNet classification.

The key insight was simple but profound: instead of programming explicit rules for recognizing objects, you could create networks of simulated neurons, show them millions of examples, and let them learn the patterns themselves. The network would adjust its internal parameters—millions of numerical weights—until it could reliably distinguish a pedestrian from a mailbox, a cyclist from a parked motorcycle, a traffic cone from a construction barrel.

This approach, called deep learning, transformed computer vision from brittle rule-based systems into flexible pattern recognizers. By 2015, deep neural networks were outperforming traditional computer vision algorithms across virtually every benchmark.

For self-driving cars, this was transformative. Early systems struggled to identify pedestrians in unusual clothing or poses. Neural networks trained on millions of images could recognize humans in thousands of different configurations—running, standing, crouching, pushing strollers, wearing backpacks, holding umbrellas. But perception was only half the problem.

The Prediction and Planning Challenge

Seeing the world is one thing. Deciding what to do about it is another.

Consider a four-way intersection where you have the right of way, but another car is approaching from the left at a speed that suggests the driver might not stop. Do you proceed? Slow down? Stop completely? The decision requires predicting what other agents will do—modeling their intentions, their awareness, their likely actions.

Human drivers make these predictions unconsciously, drawing on intuition built from experience. We notice the subtle deceleration that suggests someone is about to stop. We recognize the distracted driver who's looking at their phone. We anticipate that the car with its right turn signal might cut us off.

For autonomous vehicles, this requires what researchers call "behavior prediction"—modeling the future trajectories of every agent in the scene. Modern systems don't just track objects; they predict what each object will do over the next several seconds, generating probability distributions for different possible futures.

Waymo's sixth-generation autonomous driving system, deployed starting in 2024 and refined through 2025, processes sensor data from 13 cameras, 4 lidars, and 6 radars through integrated end-to-end neural networks. The perception module creates detailed 3D semantic maps with pixel-level scene classification for objects like vehicles and pedestrians out to over 300 meters, while the prediction system forecasts multi-modal trajectories for surrounding agents. The planning layer then evaluates thousands of candidate paths per second, selecting the one that optimizes for safety, passenger comfort (minimizing jerk and acceleration), and route efficiency under real-world uncertainty.

The computational requirements are staggering. Waymo's current vehicles contain custom-designed AI accelerators—specialized processors optimized for the matrix multiplications that dominate neural network calculations. These chips process approximately 300 trillion operations per second.

The Data Moat

But hardware and algorithms alone don't explain Waymo's dominance. The real competitive advantage is data—specifically, the scale and diversity of real-world driving experience.

By March 2025, Waymo had accumulated over 50 million fully autonomous miles. This wasn't 50 million miles of human-driven data collection. These were miles driven by vehicles making every decision autonomously, with human safety drivers present only as backup.

The Geographic Expansion

In early autonomous vehicle development, companies chose Phoenix, Arizona for testing because the conditions were ideal: clear skies, wide streets, grid layouts, minimal pedestrians. The environment was forgiving.

By 2025, Waymo had moved far beyond these training wheels. The San Francisco expansion, launched in 2024, was the acid test. San Francisco has some of the most challenging urban driving conditions in North America: steep hills, dense fog, narrow Victorian streets, aggressive human drivers, high pedestrian and cyclist traffic, complex multi-way intersections, double-parked delivery vehicles, and frequent street closures for construction or events.

In San Francisco, Waymo's vehicles have logged millions of driverless miles through fog, protests, and complex urban conditions, with reported collision rates several times lower than human drivers according to California DMV data.

The Economic Tipping Point

By mid-2025, the economics of robotaxis had crossed a critical threshold. Waymo's cost per mile had dropped to approximately $1.80, down from over $4 per mile in 2020. This included vehicle depreciation, maintenance, insurance, cleaning, charging/fueling, and the remote fleet management staff who monitored vehicles and provided assistance when needed.

For context, the fully loaded cost of personal car ownership averages ~$0.65 per mile, while Uber/Lyft rides typically cost $2–$3 per mile after driver compensation. The implications were staggering. Americans drive approximately 3.2 trillion miles annually. If robotaxis could capture even 20% of urban miles—a conservative estimate for trips that don't require vehicle ownership—that represented a $200 billion annual market.

A Parallel Revolution: Robots That Learn by Watching

While Waymo taught machines to navigate streets, another group of researchers was teaching robots to navigate warehouses, kitchens, and factory floors—and they were learning exponentially faster.

Traditional industrial robots are magnificent at repetition but helpless at adaptation. A robot arm in an automobile factory performs the same welding operation millions of times with micron precision. But if you ask it to pick up a slightly different part, or respond to unexpected variation, it fails. This brittleness made robots economically viable only for high-volume, standardized tasks. General-purpose robotics—machines that could learn new tasks quickly and handle variation gracefully—remained elusive.

The breakthrough came from applying the same deep learning revolution that transformed computer vision to the problem of robotic manipulation.

In 2023, researchers at Google DeepMind published "RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control," demonstrating that robots could learn manipulation skills by training on combinations of robot interaction data and general internet images and text. The system, called Robotics Transformer 2, could understand commands like "pick up the bag of chips" and execute them without task-specific programming. But the real revolution was happening in simulation.

The Sim-to-Real Transfer Revolution

The problem with training robots in the real world is simple: robots are slow, expensive, and fragile. If you want a robot to learn to pick up oddly shaped objects, you need thousands of attempts. In the physical world, this takes weeks. The robot tries, fails, adjusts, tries again. Engineers supervise. Objects must be reset. Broken parts get replaced.

What if you could practice in a virtual world where physics worked the same way but time ran 10,000 times faster?

This concept—called sim-to-real transfer—had been studied for decades but rarely worked. Simulated physics were too perfect, too simplified. Skills learned in virtual environments failed when transferred to messy reality. The simulation-to-reality gap was too wide. The breakthrough came from a counterintuitive insight: make your simulations deliberately imperfect.

In 2019, researchers at OpenAI showed this with Dactyl, a system that learned to manipulate a Rubik’s Cube almost entirely in simulation. The trick was ‘domain randomization’: intentionally varying physical and visual parameters—friction, object shape and mass, lighting, sensor noise, even small perturbations to gravity—so the simulator was deliberately imperfect. Policies that worked across all these randomized worlds transferred surprisingly well to the real robot hand, with no task-specific real-world practice.

By training in thousands of slightly wrong simulated worlds, the robot learned strategies robust enough to work in the real world, where conditions never perfectly matched any individual simulation but fell within the distribution of variations it had experienced. By 2025, this approach had been refined and industrialized.

Generative AI Meets Physical Manipulation

The latest breakthrough combined sim-to-real transfer with foundation models—large neural networks trained on vast datasets that develop general-purpose capabilities. By the mid-2020s, robotics companies such as Physical Intelligence, Covariant, and Tesla had begun using large language models as high-level ‘brains’ for robots. In lab and early production systems, you could specify tasks like ‘find boxes with red labels and stack them on the left pallet,’ and the model would break this into subtasks—querying vision modules to detect red labels, invoking grasping skills to pick up boxes, and calling motion planners to stack them efficiently—rather than relying on hand-crafted, task-specific code.

The robots didn't need task-specific programming. They used general-purpose visual understanding learned from billions of internet images combined with manipulation skills learned from millions of simulated and real-world interactions.

By 2025, Tesla’s Optimus humanoid robot was being trained using human demonstrations and large-scale simulation. Tesla has shown prototypes learning simple manipulation skills—such as sorting objects or basic cloth handling—from motion-captured human performances, then refining those skills in simulation (with randomized conditions) before attempting them on the real robot. The company’s stated aim is that, over time, Optimus should be able to pick up new tasks from only a small number of human demos, then generalize them via massive simulated practice.

What previously required weeks of programming could now be accomplished in hours. By 2025, these fast-learning robots were deployed in over 300 warehouses globally. By the mid-2020s, Amazon had begun piloting more versatile bipedal and mobile manipulation robots alongside its conveyor systems and fixed-path robots, exploring whether robots that share human spaces—aisles, racks, and some walkable areas—could reduce the need for highly specialized infrastructure in future warehouses.

The Manufacturing Challenge

Despite the technical breakthroughs, scaling remained expensive and complex.

Even after years of cost reduction, each Waymo vehicle in 2025 still carried tens of thousands of dollars in sensors and custom computing hardware—primarily high-end lidars, radars, cameras, and AI accelerators—representing a major ongoing challenge to scaling autonomous vehicle economics.

For general-purpose robots, manufacturing was becoming commoditized but remained capital-intensive. Producing a humanoid robot at scale required advanced manufacturing capabilities: precision machining for joints, sophisticated motor controllers, integrated sensor systems, battery packs with sufficient power density. By late 2025, Chinese manufacturers like Unitree were bringing humanoid robot costs under $50,000 at volume, making them viable for industrial and commercial pilots even if consumer household adoption remained years away.

2025: The Inflection Point

By the end of 2025 autonomous vehicles transitioned from impressive technology demonstrations to unremarkable daily transportation for hundreds of thousands of people.

The technology's immediate future is geographic expansion and domain extension. Proven capability in dense urban environments will extend to suburban and rural contexts—though these present different challenges around rare edge cases and road conditions. Trucking automation will gradually advance, though regulatory and safety hurdles remain more significant than for urban robotaxis.

For robots, humanoid form factors will proliferate beyond warehouses into retail, healthcare, and eventually residential settings. The economics continue improving as manufacturing scales. The capabilities expand as training datasets grow and algorithms refine.

But the deeper story is about machines finally gaining physical competence that matches their computational competence. For seventy years, computers excelled at abstract information processing but remained helpless in the physical world. They could calculate missile trajectories but couldn't tie shoes. They could beat chess grandmasters but couldn't safely cross a street. That asymmetry is ending. The same techniques that gave machines superhuman capabilities in narrow domains—chess, protein folding, language translation—are now giving them human-level capabilities in general physical interaction.

The implications extend beyond transportation and warehousing. Medicine will see surgical robots that learn techniques from observing procedures (already starting to happen and will only accelerate over the next few years). Construction will use robots that adapt to site-specific conditions. Agriculture will deploy machines that handle irregular crops and terrain. Disaster response will use robots that navigate rubble and unpredictable environments. The hardware costs are declining. The software is improving exponentially. The data flywheel is accelerating.

The Deeper Pattern

Step back from the specific technologies—the lidar arrays, the neural networks, the simulation frameworks—and a pattern emerges.

The twentieth century was defined by machines that amplified human physical power. Bulldozers, airplanes, power tools, assembly lines—these technologies let humans move earth, fly, build, and manufacture at scales impossible with muscle alone. But they were always tools wielded by humans. The human remained essential for perception, decision-making, and adaptation.

The twenty-first century is being defined by machines that replicate human cognitive capabilities—not just calculation but perception, prediction, and learning.

Waymo's vehicles aren't assisted driving systems that help human drivers. They are complete replacements that render human drivers unnecessary. Warehouse robots aren't tools that augment human workers. They are substitutes that perform the same tasks with comparable or superior efficiency. This is different. This is machines gaining competence in domains that, until very recently, seemed fundamentally human.

The technology remains young. The safety isn't perfect. The economics aren't universally favorable. The social disruption will be substantial and uneven. But there is no doubt that the machines are learning. And they're becoming a big part of our everyday lives.

Chapter 7: The Agentic Web (AI That "Does")

Noor Shaker

The Breakthrough That Changed Everything

For three years, from late 2022 through 2024, artificial intelligence meant conversation. ChatGPT arrived in November 2022 and within months became the fastest-growing consumer application in history, reaching 100 million users by January 2023. GPT-4, Claude, and Gemini followed, each more capable than the last at generating text, answering questions, and holding discussions that felt remarkably human.

But they all shared fundamental limitations. They stopped after each response, requiring constant human prompting to continue. They had no persistent memory—each conversation started fresh, with no recall of previous interactions. They hallucinated facts with unsettling confidence, inventing citations and statistics that sounded authoritative but were entirely fictional. And crucially, they couldn't act. Ask ChatGPT to book a flight, and it would tell you how to book a flight. Ask it to resolve a supply chain crisis, and it would outline a solution. But in every case, execution remained entirely in human hands.

This created what researchers called the "execution gap"—the distance between AI suggesting a solution and someone implementing it. The technology could think but not act, advise but not execute, plan but not perform.

Throughout 2025, we've seen signs of the gap closing. A new generation of AI systems emerged that didn't just respond to prompts—they pursued goals autonomously. They broke complex tasks into steps, used tools to interact with software systems, evaluated their own work, and iterated until objectives were met. Issues such as maintaining context across extended interactions, building persistent memory of user preferences and prior work are being gradually resolved. Mechanisms to resolve critical matters, such as hallucinations and privacy, are increasingly being developed to verify information through tool use—checking databases, running code, and validating outputs against reality rather than relying solely on pattern-matched training data.

The terminology shifted to match the capability. These weren't "chatbots" or "language models"—they were "agents." And by December 2025, agents were handling tasks that six months earlier had seemed impossible. GitHub Copilot Workspace and similar tools now accept feature requests in plain English and generate complete implementations—writing code, creating tests, running them, debugging failures, and submitting pull requests for human review. Salesforce's Agentforce and ServiceNow's AI agents manage customer inquiries from detection through resolution, pulling customer history and updating records across multiple systems without human intervention at each step.

The shift from chatbot to agent represents more than incremental improvement. It is the difference between a reference book and a colleague. It is a new type of operating system that will redefine how we interact with the digital world.

The Move to Small Language Model

While agentic systems transformed what AI could do, a parallel revolution transformed where it could run. The surprise of 2025 wasn't just smarter AI—it was smaller AI that performed nearly as well.

From 2020 to 2024, AI development followed a simple rule: bigger is better. GPT-3 had 175 billion parameters. GPT-4 reportedly exceeded one trillion. These models required massive data centers, cost hundreds of millions to train, and consumed megawatts running inference. A single query to GPT-4 in 2023 used roughly the same energy as charging a smartphone for 24 minutes—staggering at global scale.

This created an obvious constraint: powerful AI was limited to capital-intensive companies, subject to latency delays, and raising privacy concerns as sensitive data travelled to remote servers for processing. For applications requiring instant response, offline operation, or absolute data privacy, public AI was inadequate.

With the increasing interest in running LLMs locally and privately, and the long-lasting trend of Moore’s law, LLM models of 2025 are smaller, more cost-efficient and almost as powerful as large ones. They compressed GPT-4-level performance into models 100 to 1,000 times smaller. Microsoft's Phi-4, announced in December 2024, packed 14 billion parameters while matching larger models on mathematical reasoning. Google's Gemini Nano, available in variants of 1.8 billion and 3.25 billion parameters, ran on smartphones while handling complex queries. DeepSeek offers models in various sizes (like 1.3B, 7B), open access and can be run on a local laptop.

The key innovation wasn't a single technique but a combination of approaches. Instead of training on the entire internet, Small Language Models (SLMs) trained on carefully curated, high-quality datasets focused on specific domains. A medical SLM might train exclusively on peer-reviewed papers, textbooks, and clinical notes—10 million carefully selected documents rather than 10 billion random web pages. This improved signal-to-noise ratio, allowing smaller models to achieve expert-level performance in narrow domains.

Technical optimizations added further compression. Quantization reduced neural networks from 32-bit floating-point numbers to 8-bit or even 4-bit integers with minimal accuracy loss, cutting memory requirements by 75-90%. Pruning identified and removed redundant neural connections, like trimming dead branches from a tree. Mixture of Experts (MoE) architecture, routes each input to a small subset of specialized "expert" sub-networks rather than activating the entire model, drastically reducing computation while maintaining capability.

The Death of Search (As We Knew It)

Perhaps the most visible transformation of 2025 was one most people didn't consciously notice: the quiet death of traditional search engines.

For twenty-five years, web search followed a simple pattern. You typed a query. Google (or Bing, or DuckDuckGo) returned a page of links—typically ten, hence "ten blue links"—ranked by relevance. You clicked, read, evaluated, and often reformulated your query, repeating the process until you found answers. The model had become so ingrained that "googling" became a verb. Entire industries—SEO, content farms, affiliate marketing—existed to manipulate rankings on these search result pages.

In early 2025, Google, Microsoft, and Apple completed a transition that began in 2024: search engines stopped returning lists of links as the primary result. Instead, they generated direct answers. Ask "What are the health effects of intermittent fasting?" and rather than seeing links to Mayo Clinic, WebMD, and various diet blogs, you receive an AI-generated text summarizing current scientific consensus, noting areas of uncertainty, citing specific studies inline, and offering an organized overview of benefits, risks, and considerations. Links appear below in the traditional way.

This transformation required solving multiple technical challenges. Generative AI models have knowledge cutoffs—dates after which their training data ends. To provide current information, 2025 search engines combined large language models with live web indexing through a process called Retrieval-Augmented Generation (RAG), originally introduced in a 2020 paper by researchers at Meta AI Research. When you query, the system generates a search plan identifying which sources might contain relevant information, retrieves current web pages and databases, extracts relevant passages, feeds this context to the LLM alongside your query, and generates a synthesis with inline citations. This process was refined throughout 2024-2025 to operate in under two seconds—comparable to traditional search.

Unlike traditional search where users evaluated source credibility themselves, generative search made the AI responsible for information quality. Systems implemented multi-layered verification: source reputation scoring that prefers established medical journals over random blogs for health queries, cross-referencing that requires multiple independent sources to confirm contested facts, uncertainty quantification that explicitly states when information is disputed, and primary source preference that favors original research over aggregators.

The shift to generative search created immediate economic fallout. If users get answers directly from search engines, they don't click through to websites. Analytics firms reported that many content sites saw traffic decline substantially from March to December 2025. Entire business models built on SEO-optimized content and ad revenue faced existential crisis. The incentive structure flipped. Rather than churning out hundreds of shallow articles to capture search traffic, publishers invested in unique, authoritative content that AI systems would cite.

The Integration Moment

What made 2025 remarkable wasn't just individual technologies—agents, SLMs, generative search—but their convergence. Consider a realistic scenario from October 2025: A manufacturing plant manager asks their AI assistant, "Why is Line 3 production down 8% this week?"

The system searches internal databases and real-time sensor data, using generative search techniques to synthesize information from multiple sources. It identifies three anomalies: a supplier delayed shipments, a robotic arm is miscalibrated, and ambient temperature exceeded optimal range. Displaying agentic behavior, it plans corrective actions and executes autonomously where possible, adjusting robotic calibration through API, ordering replacement parts using the procurement system, and scheduling HVAC maintenance. For decisions requiring human judgment, it consults the manager: "Supplier offers 15% discount for accepting further delays or we can source from backup supplier at 8% premium—which do you prefer?" The entire analysis runs locally on edge servers for response time under 200 milliseconds and data privacy compliance, leveraging small language model efficiency.

This closed-loop cycle—question, analysis, action, verification—is happening across manufacturing, healthcare, logistics, finance, and creative industries. The machinery was invisible to most users. They simply noticed that technology had become more helpful, more proactive, less demanding of their attention and effort.

The agentic revolution arrived with legitimate concerns. Unlike previous AI waves that augmented human work, agents began replacing it. Roles involving routine digital tasks—data entry, simple customer service, basic coding—saw significant automation. Estimates of job displacement varied widely, but it is clear that AI is moving from assistant to substitute for an expanding range of work.

Security risks emerged as agents gained write-access to systems. "Prompt injection" attacks—where malicious actors embedded hidden instructions in emails, published articles or web pages that agents processed—became a serious threat. As agents became more autonomous, ensuring they reliably pursued intended goals—and didn't pursue unintended harmful goals—moved from theoretical concern to practical urgency.

Why 2025 Was the Inflection Point

Several convergent factors made 2025 the year of agents. Hardware advances—NVIDIA H200 GPUs, Google TPU v5, specialized AI accelerators—dropped inference costs substantially from 2023, making continuous agent operation economically viable. Algorithmic maturity moved techniques like chain-of-thought prompting, tool-use protocols, and verification systems from research papers to production. Standardized APIs and open access models allowed agents to interact with thousands of services without custom integration.

Perhaps most importantly, public readiness had matured. Consumer experience with ChatGPT and similar systems from 2023-2024 built familiarity and changed how we are used to interact with digital contents and do tasks. The idea of "AI that does things" seemed like natural evolution.

Looking Forward

As large language models and agentic AI continue evolving and become embedded in our daily interactions, we are coming to better appreciate where and how they serve us best and what their limitations are. LLM systems, while very powerful, still hallucinate and are far from being trustworthy. The relentless march of Moore's law means these tools will eventually become commoditized, mostly in the form of local personalized models running on edge devices. Commercial models for many existing companies will be challenged as capabilities that once required cloud infrastructure migrate to smartphones and laptops.

Agentic AI is still far from its glory days, with useful applications being discovered and strengths and weaknesses being explored. Co-coding with AI agents has emerged as one of the prime applications and has proven remarkably useful, measured by the explosion of apps and websites now largely implemented by AI agents in late 2025. But code quality, structure, and optimization are still far from optimal. These systems operate at roughly a junior software engineering level, requiring an expert human in the loop to guide architectural decisions and review output before deployment.

The most widely adopted and immediately valuable applications for agentic AI are automating tasks that require manual execution rather than deep expertise—reformatting or reorganizing data across systems, improving or creating content at scale, automating bulk communications, managing CRM workflows, and similar operational overhead. These tasks previously consumed substantial human time and salary expense. Agentic AI eliminates this overhead almost entirely, reducing both time and cost by orders of magnitude. Dismissing this as mere efficiency improvement misses the point. When a single agent can perform in minutes what once required hours of human labor across an entire team, the accumulated impact reshapes how organizations allocate human capital. The transformation is fundamental, even if the individual tasks are mundane.

More demanding applications—scientific discovery, complex applications, tasks that require genuine expertise—are still in early progress. These models will certainly become more capable as researchers continue refining and fine-tuning them on better, task-specific data. But this journey will likely take several years, not months. Integration remains a significant bottleneck, with these new systems needing to interact with legacy frameworks and datasets built for human operators.

Regulatory systems also need to catch up to the pace of innovation, especially when it comes to critical applications like healthcare. Having agentic systems that automate insurance claims, streamline operations, take notes, and summarize documents is valuable. But what would be truly transformative is having a connected system of multiple agents working together—diagnostic support, trial enrollment, personalized treatment, real-time monitoring—all coordinated seamlessly. This requires not only advancement on the research side (which is mostly ongoing) but also rapid evolution from regulatory bodies to ensure thorough testing and safe implementation of these systems at the pace innovation demands.

The next few years will witness something no less than revolutionary. From AI systems conducting scientific discoveries to quiet but certain transformations of our daily work routines, from robotic systems planning and executing complex tasks autonomously to multi-agent coordination on goals that once required dozens of human specialists—the frontier is expanding rapidly. It is, without question, a remarkable decade to be living through.