Toward a Science of Intelligence: Unifying Physics, Neuroscience & AI

Surya Ganguli

Artificial intelligence stands poised to transform our society, yet we hardly understand how it works. A synthesis of physics, neuroscience, and AI can fulfill an urgent need: to build a new, unified science of intelligence that explains and improves how intelligence emerges across both artificial and biological neural networks. I discuss four ways this synthesis has begun to and will continue to unfold. First, powerful analytic tools from the physics of complex systems will provide insight into how large neural networks learn and compute. Second, neuroscience will provide clues into bridging the many orders of magnitude advantages that biological intelligence retains over AI. Third, we can go beyond evolution to instantiate neural algorithms in quantum hardware, leading to new devices through AI-led codesign of physics and computation. Fourth, we can meld minds and machines by building digital twins of the brain, yielding insights into not only intelligence but also consciousness and the sense of self through causal modeling and control. Overall, AI will shed light on the nature of our physical and mental realities, raising profound questions about the role of human understanding in the age of AI.

The first glimmers of human-like intelligence appeared in Africa a few million years ago, culminating in the brain of our species *Homo sapiens* about three hundred thousand years ago. Our ancient ancestors no doubt peered both out into the night sky to contemplate the nature of physical reality and inward to ponder the nature of their own mental reality.

In the last century, we have developed a deep understanding of physical reality through precise mathematical laws governing the behavior of space, time, matter, and energy. But we are only beginning to understand our mental reality. How does human intelligence and consciousness emerge from one hundred billion neurons connected by one hundred trillion synapses? Modern neuroscience has laid strong foundations for attacking this grand question. But when it comes to our own mental capabilities, we seek not only to understand but also to recreate these capabilities in machines, sometimes fashioned after our own image. In essence, humans, as products of evolution, yearn to play the role of creator. This yearning permeates literature, ranging from Mary Shelley's *Frankenstein* to Isaac Asimov's *I, Robot*.

https://doi.org/10.1162/DAED.a.984

In the last decade, artificial intelligence has made striking progress in realizing this yearning, making highly intelligent systems capable of vision, language, reasoning, imagination, and more. These capabilities emerged in an incredibly nonintuitive and *implicit* manner from scaling up large neural networks, fit to large amounts of data, on large compute clusters. Such AI systems stand poised to transform our economy, society, and the very nature of scientific research itself. However, quite alarmingly, such AI systems possess large and mysterious surfaces of *both* capabilities *and* fragilities, and we barely have any scientific understanding of either. Thus, there is a fierce urgency to develop a deeper scientific understanding of AI to ensure the development of trustworthy, robust AI systems aligned with human needs.

There are immense opportunities for advancing our scientific understanding of intelligence, across both brains and machines, through an interdisciplinary synthesis of physics, neuroscience, and AI. First, physics has developed powerful theoretical tools for the analysis of complex systems consisting of many billions of interacting particles with intricate, dynamic laws that influence their collective behavior and give rise to surprising "emergent" properties: new characteristics that arise from the interactions but are not present in the individual particles themselves. Similarly, learning in neural networks involves many billions of interacting weights, with complex learning dynamics driven by data, giving rise to surprising emergent capabilities. Thus, ideas from physics have and will continue to have a powerful impact on AI. Second, neuroscience and the allied fields of psychology and cognitive science reveal multiple orders of magnitude advantages that human intelligence still retains over AI, especially in terms of data efficiency, energy efficiency, and robustness. Neuroscience in this way provides both beacons and clues toward improving AI. Third, one can combine neuroscience and physics to go beyond evolution, to instantiate the neural algorithms for computation discovered by evolution but through quantum hardware that evolution could not discover. This yields new types of quantum neuromorphic technologies inspired by the synthesis of physics and neuroscience. Fourth, AI can advance our understanding of biological intelligence and consciousness by turning back toward us and employing bidirectional feedback loops that meld brains and machines. Thus, while human intelligence gave birth to artificial intelligence, our child will become our teacher and will help us peer into the nature of our own mental reality as well as physical reality itself.

The rest of the twenty-first century will be an exciting intellectual adventure, in which physics, neuroscience, and AI work hand in hand to teach each other and us about the science of intelligence.

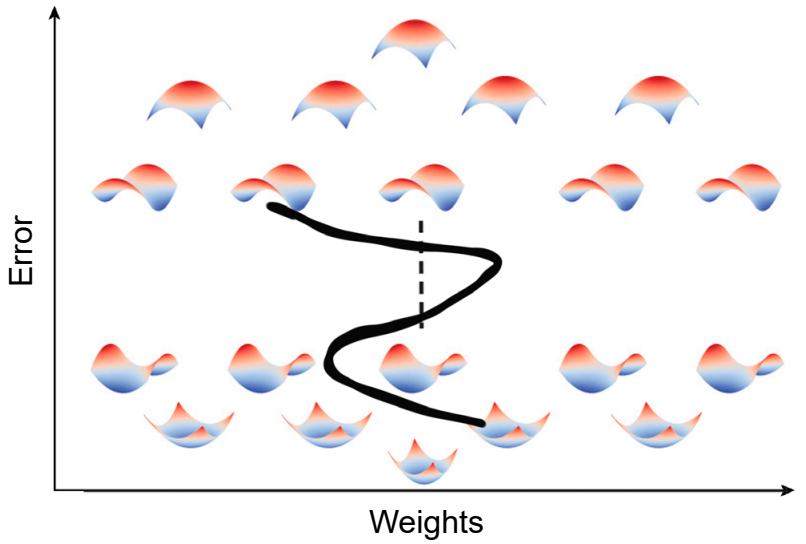
Physics has long shaped the development of artificial intelligence. While so-called fundamental physics historically strived to understand nature at the smallest subatomic scales, condensed matter physicist Phillip Anderson argued in his influential 1972 essay "More Is Different" that an equally fundamental

problem involves understanding how surprising and nonintuitive properties at a larger scale *emerge* from the interactions of many simple elements at a smaller scale.[1] For example, how does the "wetness" of water emerge from the interaction of many H_2O molecules? Why do metals conduct? How does glass achieve its simultaneous rigidity and fragility? The field of condensed matter physics arose to develop powerful theoretical tools to analyze complex matter consisting of many interacting atoms.

However, the notion of emergence extends beyond physics. For example, how does cellular life emerge from the interactions of many nonliving molecules? How does intelligence emerge from the interaction of many nonintelligent neurons? Bolstered by their success in condensed matter physics, many physicists turned to biology and neuroscience. John Hopfield was one such physicist: In 1982, he developed a breakthrough model of associative memory through an analogy to magnetic spin systems in physics.[2] This model explained how networks of neurons could store new facts and recall facts given partial information. In turn, the Hopfield model led to the Boltzmann machine and deep-belief networks, both precursors to modern deep learning.[3] For this arc of discovery, John Hopfield and Geoffrey Hinton won the 2024 Nobel Prize in Physics. And this arc continues: the modern transformer architecture is closely related to a modified Hopfield network.[4]

Another influential line of work in the theoretical physics of complex systems is that of Giorgio Parisi, who won part of the 2021 Nobel Prize in Physics. Parisi and colleagues showed how to analyze the geometry of high-dimensional energy landscapes over many particles and describe how the particles move in this energy landscape geometry.[5] We are used to thinking about functions over two-dimensional space (for instance, the height of an actual hilly landscape on the surface of the earth). But the energy landscape of a physical system is a function of the position of many billions of particles. Similarly, the loss landscape of a neural network is a function of the values of many billions of parameters. Thus, there can be commonalities in how cooled physical systems lower their energy as particle positions change and in how neural networks learn to lower their loss as their weights change. Indeed, some features of the geometry of and dynamics within high-dimensional landscapes can exhibit universal properties. For example, why doesn't neural network learning get stuck in local minima with high error, like a ball rolling down a rugged hilly landscape might get stuck in a high valley? The key idea is that in high but not low dimensions, such high-energy minima are exponentially rare. Intuitively, suppose you are at an energy or loss minimum in a billion-dimensional space. What are the chances that motion in *all* one billion directions will go up? The answer is extremely unlikely, unless you are *already* near the bottom of the landscape. Local minima therefore typically only occur in a range of energies or losses close to that of the global minimum. Instead, saddle points proliferate, which always provide a way down (Figure 1). This intuition may play a key role in explaining why neural networks can often achieve low training error at all.[6]

Figure 1
Learning Dynamics as Descending Motion in a High-Dimensional Error Landscape



Learning dynamics (black curve) follow a descending motion in a high-dimensional error landscape over many weights. This landscape is riddled with many maxima, saddle points, and minima, but the minima typically occur only near the bottom. Source : A modified form of Figure 2(a) in Yasaman Bahri, Jonathan Kadmon, Jeffrey Pennington, et al., "Statistical Mechanics of Deep Learning," *Annual Review of Condensed Matter Physics* 11 (2020), https://doi.org/10.1146/annurev-conmatphys-031119-050745.

Physical thinking has also played a key role in more recent advances in deep learning.[7] For example, diffusion models – which now power much of generative AI, creating new images, text, proteins, and more – were first inspired by attempts to violate the second law of thermodynamics.[8] This law asserts that in closed systems, ordered patterns will overwhelmingly evolve into disordered patterns, thereby increasing the entropy (disorder) of the system. The key idea behind diffusion models is to replace the laws of physics with a neural network trained specifically to *decrease* entropy. Once trained, the neural network, unlike the laws of physics, can turn disordered patterns into structure. More recent theoretical work has even been able to quantitatively explain the origins of creativity in such models.[9]

Another influential finding guided by physical thinking was the discovery of scaling laws, or predictable power-law relationships between improved perfor-

mance and resources, like network size, data size, and training compute.[10] Such scaling laws have captured the imagination of industry and motivated significant societal investments in energy, compute, and data collection. And a theoretical explanation of data scaling laws for language models has emerged only recently.[11] Though as we shall see in the next section, it should be possible to achieve much more efficient scaling.[12]

In summary, AI systems are fundamentally large and complex interacting systems, much like many other physical systems; and techniques from physics have and will continue to play a key role in the further analysis and development of AI. In particular, by building on the success of physics, I believe we will be able to develop new mathematical ideas for the analysis, development, and simplification of nonlinear maps between high-dimensional spaces. Such maps are ubiquitous in all of AI: for example, maps from inputs to outputs in language models and diffusion models, and maps from data to weights in neural learning processes. Indeed, many of the mysteries of how neural networks learn and compute lie in our current inability to understand such nonlinear maps between high-dimensional spaces. But the success of physics in understanding the interplay between nonlinearity, stochasticity, and high dimensionality in other settings provides a fertile springboard for generating a new mathematical framework for comprehending and shaping high-dimensional nonlinear maps in AI, thereby demystifying and empowering learning and computation in neural networks.

Neuroscience and the allied fields of psychology and cognitive science stand as taunting sirens, reminding us that biological intelligence still outperforms AI by many orders of magnitude along several axes, including data efficiency, energy efficiency, and robustness. First, data efficiency: Modern large language models (LLMs) are trained on the order of ten trillion words. However, humans are exposed to only about one hundred million words – five orders of magnitude less. It would take us about 240,000 years to read everything an LLM read. So, at least in terms of language experience, we are incredibly efficient learners. There are many conjectures why. Of course, language is not our only experience. We also get visual inputs and have embodied experience.[13] However, it is less clear how looking at more videos of scenes and objects could endow the collective human mind with the requisite mathematical and reasoning capabilities to pierce into the mysteries of the universe and discover novel ideas like quantum mechanics and general relativity within only ten generations, from Newton to Einstein. This remarkable "out of distribution" generalization remains uniquely human.

To understand the potential origins for data inefficiency in machines, imagine we taught our children the same way we pretrain an LLM. We would repeatedly give a child a random sentence from the internet and then tell them the next word. *Nothing more*. No child could meaningfully learn that way, making it all the more remark-

able that LLMs can learn anything at all after seeing one trillion sentences (LLMs also undergo additional and significant post-training). Instead, human learning is a multifaceted, multi-timescale process in which we carefully sequence information given to our children. Moreover, children, like scientists, learn actively, performing experiments and choosing their own training data. As we grow older, we form useful abstractions, discard useless ones, and modify them in one shot in response to new facts. And over generations we transmit the most useful information in treasured books, signaling to our descendants what is worth learning. Achieving the five orders of magnitude data gap between brains and machines may well require many elements of this rich human teaching process. We must go beyond machine learning to develop a new science of machine *teaching*.

Consider next the axis of energy efficiency. Brains consume a miniscule amount of power: 20 watts. For reference, an old incandescent light bulb might consume 100 watts. In a sense, then, we are all dimmer than light bulbs. In contrast, training an LLM can consume tens of megawatts, about six orders of magnitude more. Moreover, there is talk of building nuclear power plants to power gigawatt data centers. So, what went wrong? After all, evolution did not need to build nuclear power plants before endowing humans with the intelligence to invent them. Unfortunately, the fault of high power consumption in machines may lie in the very choice of digital computation itself. To compute, we rely on extremely fast and reliable bit flips, implemented through the flow of many electrons in energy-consuming transistor circuits. However, the laws of thermodynamics demand an energy cost for every fast and reliable bit flip. Biological computation took a very different route. Every intermediate step of a biological computation is slow and unreliable. Molecules diffuse. Synapses fail. The clock period of neurons is milliseconds, not nanoseconds, as in GPUs. However, the composite computation is just good and fast enough for survival. In essence, biology does not rev its engine any more than it needs to to get its final computations right. The sloppy, intermediate unreliability of biology is a feature, not a bug – a feature of incredibly energy-efficient design.

To close the six orders of magnitude energy gap between brains and machines, it may be necessary to rethink our entire technology stack, from electrons to algorithms, and go back to the drawing board to develop new theories of reliable computation with stochastic, analog devices. Indeed, the field of stochastic thermodynamics has developed powerful tools to analyze tradeoffs between energy, accuracy, and speed of stochastic processes that could underlie computation.[14] For example, researchers have recently derived a Pareto optimal frontier of the minimum error achieved by molecular sensors as a function of their energy budget, and revealed the structure of optimal families of sensors that trace this frontier.[15] This work shows that molecular sensors *must* spend more energy to perform more accurate sensing. Moreover, the structure of optimal sensors is reminiscent of G-protein coupled receptors, which reside inside every neuron to sense external signals. At a

more global level, neuroscience has now developed novel sensors to measure not only neural activity but also ATP (adenosine triphosphate) consumption. ATP is the chemical fuel that powers all cellular activity. Researchers have recently performed simultaneous measurements of neural activity and ATP consumption across the entire *Drosophila* brain and found evidence for a predictive energy-allocation hypothesis.[16] Essentially, the brain appears to behave like a smart energy grid: it predicts where, when, and how long energy will be needed, and it delivers just the right amount of energy, at the right location, at the right time, to power predicted future neural activity. In short, five hundred million years of vertebrate brain evolution has had ample time to optimize coupled energy and information flows across the brain, and it is likely that there are many useful secrets we can learn from studying such flows.

Finally, the last axis of robustness yields a major gap. AI still makes egregious errors that no human would make. The classic version of this is adversarial examples: tiny image perturbations that are imperceptible to a human but completely fool an image classifier.[17] This problem still has not been solved. Even modern LLMs are subject to adversarial examples.[18] Relatedly, image classification networks confidently perceive exponentially many images as belonging to any given semantic concept (for instance, cat or dog), yet these images look like utterly unrecognizable noise to humans.[19] So overall, the way that AI perceives and thinks is fundamentally different, and less robust, than the way we do. In the famous words of Kendrick Lamar, "they not like us."

In summary, while AI has many different, complementary, and superhuman strengths in terms of the breadth of capabilities, we still have a lot to learn from biological intelligence in terms of data efficiency, energy efficiency, and robustness. I envision a future in which we go from megawatts to watts through alternative physical substrates for computation that radically depart from current digital technology, by embracing slow, stochastic, analog physical primitives that nevertheless yield emergent reliability at a systems level through immense parallelism. I also envision radically more data-efficient AI that embraces sound principles of learning gleaned from child development, including active learning, curriculum design, social learning, and cultural evolution. Such radical data efficiency may also require embracing sound principles of developmental neurobiology, which has led to a conserved modular brain architecture spanning five hundred million years of vertebrate brain evolution, including the cortex, basal ganglia, thalamus, and cerebellum, each operating at different levels of hierarchical control, each with its own learning rules spanning unsupervised, supervised, and reinforcement learning, and each communicating with the others through specific bandwidth-limited communication protocols. New microscopes enable us to record from all these regions at once and glean how such heterogeneous, hierarchical, and modular systems empower efficient learning, compared with data *inefficient* end-to-end training of a single

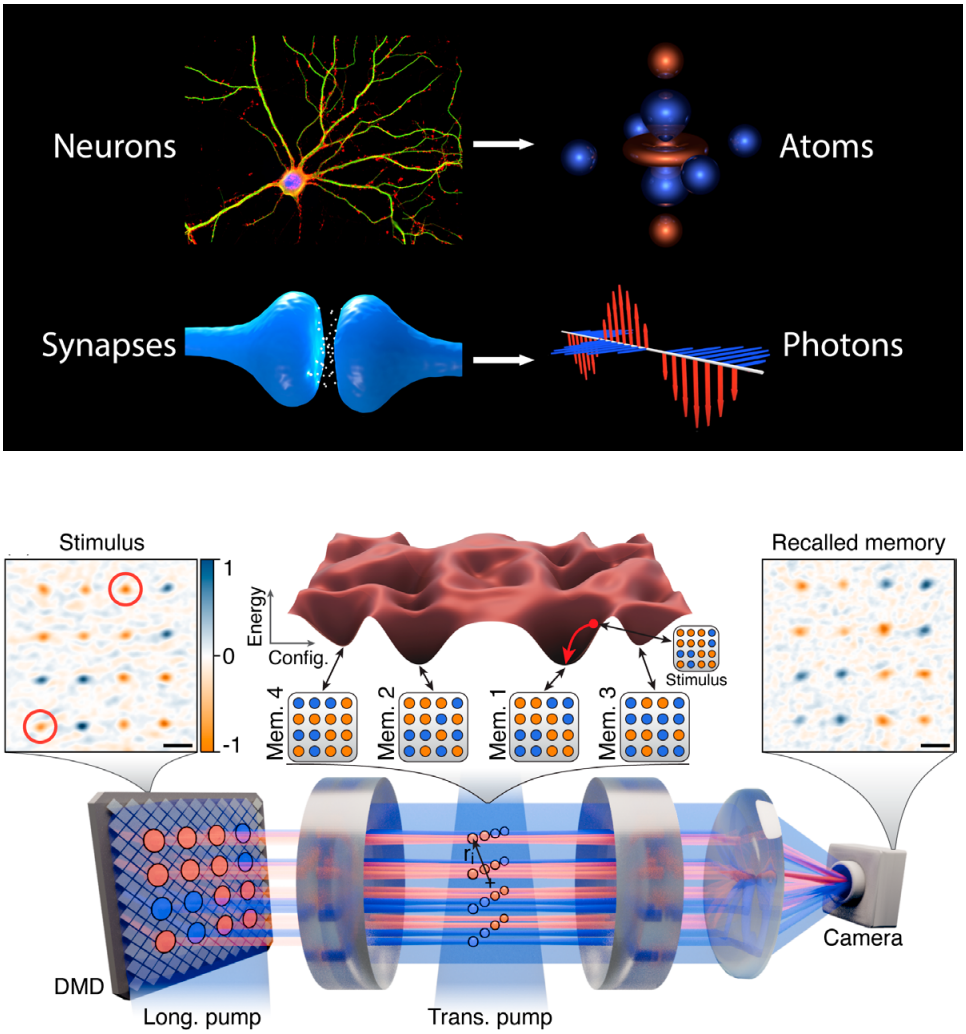
homogeneous monolithic architecture like the transformer.[20] Finally, I envision much more robust neural representations and computations arising from these energy-efficient and data-efficient learning systems, leading to AI that does not make egregious errors.

Evolution has created remarkably efficient and robust biological intelligence. But as in all technology development, we should not be limited by the design principles discovered by evolution. Indeed, recent work has explored implementing neural algorithms in quantum hardware. In essence, neurons are replaced by atoms, while synapses are replaced by photons (Figure 2, top). The idea is that different states of neurons (firing or not firing) are replaced by different electronic states of atoms (excited or not excited). Moreover, when an atom shifts electronic states, it can emit a photon, which travels to other atoms and changes their electronic states. By arranging such atoms and photons in a cavity quantum electrodynamics (cQED) system (Figure 2, bottom), one can realize a Hopfield associative memory, the same Hopfield model that inspired models of memory in neuroscience. However, interestingly, the dynamics of atoms and photons in this system allow for more robust memories than the original Hopfield model.[21] Moreover, quantum entanglement between different atomic spins can drive optimization to even lower energies.[22] Additionally, a different type of architecture, the coherent Ising machine, which consists of a network of interacting photons and whose classical dynamics resemble a neural network, can be used to solve hard combinatorial optimization problems by gradually changing a high-dimensional energy landscape.[23]

These alternative physical computing paradigms provide new ways to natively match the abstract dynamics of actual algorithms to the physical dynamics of atoms and photons that could naturally implement these algorithms. This is quite distinct from digital technology that abstracts everything in terms of bits and gates as primitives, with algorithms built on top of that. As a concrete case, consider addition. In digital computers, this is implemented in adder circuits consisting of many transistors. In contrast, biological neurons simply add their neighboring voltage inputs by using Maxwell's equations of electromagnetism, which express how to add voltages through the principle of superposition. Thus, biology beautifully matches this computation directly to the native physics of our universe. Similarly, cQED systems natively implement associative memory while coherent Ising machines natively implement combinatorial optimization. It is exciting that after evolution harnessed the laws of physics to discover neural computation, we can now implement such neural computation in never-before-realized quantum substrates, not by bending the laws of physics to the will of our algorithms but rather by matching our algorithms to the laws of physics themselves.

Going further, I envision a future in which AI will *codesign* computation and physics *together*. For example, to achieve a particular human specified computational

Figure 2
Quantum Neuromorphic Computing in Cavity Quantum Electrodynamics



(Top) Neural algorithms in quantum hardware use atoms in place of neurons and photons in place of synapses. Source: Figure by the author. *(Bottom)* A cavity quantum electrodynamics system capable of forming a Hopfield associative memory. Source: Brendan P. Marsh, David Atri Schuller, Yunpeng Ji, et al., "High-Capacity Associative Memory in a Quantum-Optical Spin Glass," arXiv (2025), https://doi.org/10.48550/arXiv.2509.12202.

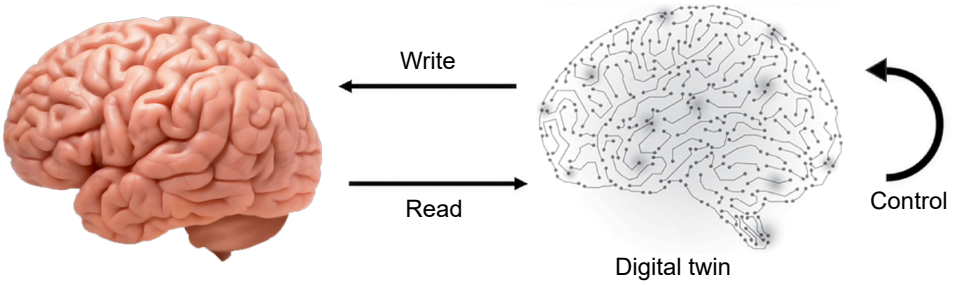
goal, AI may codesign the intermediate computational steps required to achieve this goal, with the intermediate physical states in novel physical systems (including new materials and modes of light-matter interaction) that map to these computational states. This AI-driven codesign of computation and physics could lead to special-purpose computing devices of unprecedented speed and energy efficiency. This codesign stands in stark contrast to what is done today : at one extreme we design general purpose digital technology to implement any Turing computable function, while at another extreme we examine fixed physical devices (for example, particular qubit implementations) and try to design error-corrective computations on top of them. However, AI-aided search in the incredibly large *joint* design space of computation *and* physics could yield new types of physical computers unimaginable today.

Simultaneous advances in neurotechnology and AI have made the melding of minds and machines an exciting possibility ; a path forward is now emerging, as follows (Figure 3). First, in animal neuroscience, we can record thousands of neurons linked with different behaviors. Second, AI can use these data to build digital twins of the brain : circuit models that can predict neural activity and behavior in new situations. Third, these twins allow many rapid *in silico* experiments that are impossible in animals, thereby accelerating discovery and further revealing computational principles through explainable AI. Fourth, we can use the digital twin to learn how to speak to the brain by applying control theory to the twin. This involves designing neural stimulation protocols to make the twin behave in a desired manner, and then transferring these protocols to the brain to get it to behave in the same way. Such protocols can be designed to test scientific theories, actively learn models, or even achieve desired clinical states in a principled, model-driven manner.

As an example, researchers recently built a digital twin of the retina that could predict the retinal response to videos of natural visual scenes and reproduce two decades of experiments on the retina.[24] Moreover, explainable AI applied to the twin automatically provided both conceptual insights and novel testable predictions. Deeper in the primate visual system, control theory applied to a digital twin of the visual cortex revealed stimuli that made neurons in the twin and *also* in the brain fire at high rates.[25] Earlier work built simple decoders of the visual cortical code and used these decoders to design direct stimulation patterns to the brain to make the mice behave as if they saw a particular image.[26] Remarkably, stimulating only twenty well-chosen neurons was sufficient to trigger a percept. Additionally, researchers recently built a digital twin of the epileptic brain, used explainable AI to understand how it works, derived control signals to control seizure amplitude in the twin, then successfully transferred these control signals to control seizure amplitude in the actual brain.[27]

Overall, this new bidirectional interplay between brains and machines, in which machines learn the language of the brain and speak back to it in modified ways,

Figure 3
Melding Minds and Machines



Source: Figure by the author.

either for scientific insight or clinical control, holds immense promise for understanding, augmenting, and repairing the brain. We are only beginning to build digital twins of the brain, but such increasingly expansive models spanning sensation to action could reveal principles of robust perception, motor action, learning, and cognition.[28]

In the longer term, an intriguing question lies in deeper phenomena that for now are uniquely biological and not yet found in machines: our consciousness and sense of self. How do we become aware of ourselves as agents acting in the world, and where in the brain is this sense of self rooted? Decades of neuropsychiatry provide clues through disorders of the sense of self arising from defects in various brain regions.[29] Such disorders include zombie-like epileptic states where sensorimotor acts (such as drinking from a cup) can occur without consciousness. Also, in patients with chronic pain, cingulate cortex lesions can dissociate felt pain from any associated emotional suffering. Other altered states of consciousness associated with brain damage include coma, vegetative states, locked-in syndrome, Alzheimer's, schizophrenia, blindsight, and various agnosias. Additionally, dream states and psychoactive drugs can induce dissociations of the sense of self and hallucinations. Every one of these profound alterations of the sense of self is rooted in mechanistic changes in information propagation through neural circuits. The time is now ripe to move away from studying the sense of self as pure philosophy and instead to develop an integrative theory that links behavior, neurobiology, and the principles of computation, possibly by creating system-level digital twin–like architectures of the brain.

Neuroscientist Antonio Damasio provides an intriguing hypothesis for the origin of the sense of self.[30] Every complex organism has two disparate neural sensory maps : one representing the external world through our five senses and another representing our inner world (such as internal states of muscles, visceria, hormones, and metabolism). A fundamental goal of any organism lies in homeostasis of its inner world, keeping all these internal states within safe physiological ranges. Of course, the external world can have fundamental causal influences on our inner world. The detection and imbibing of water outside can alleviate low water levels inside. The detection of a predator outside may mark the dissolution of the inner self if no action is taken. Thus, any organism that has complex sensory arrays of both the outer and inner world must develop control loops that combine information from these arrays to make decisions and take actions (such as moving toward water and away from predators). Such control loops, to act optimally, must necessarily and predictively model the causal impact the outside world can have on the inner world. It is plausible that our very sense of self, our continuous awareness that we are an entity that is both impacted by the world and can act on it, is biologically and physically rooted within causal predictive models of how the world can change our inner self, which in turn guide control loops that transform sensations of the outer world into actions that adaptively regulate our inner world.

When viewed this way, our sense of self emerges from computational requirements of causal modeling and control. To make this theory mathematically precise and test it, I believe the way forward is to build systems-level digital twins of the brain that instantiate these ideas. As a start, such twins will have components akin to different brain regions, and deficits in the twin will yield deficits in behavior akin to patients with an altered sense of self. We are very far from achieving such a twin ; it may require years of integrative work across neurobiology, psychiatry, computation, causal modeling, and control theory. But just as the machines of today achieve remarkable feats of vision and language, our machines of tomorrow may achieve a rudimentary sense of self-awareness as agents in the world, arising as an emergent property of optimally modeling the causal effects of the world on their internal state and using such causal models to decide how to act. The emergence of such machines may help us demystify the origins of self-awareness in humans and animals, just as today's machines already deepen our understanding of intelligence itself.

Events on our planet have taken a wondrous trajectory, from the development of physical systems to evolution to biological intelligence to human engineering and eventually to artificial intelligence. In each step, new phenomena emerge that generate new understandings of earlier phenomena. Evolution, operating on the laws of physics, gave rise to biological intelligence, culminating in the human mind. The human mind then turned back to develop a precise understanding of physical laws. It then bent physical laws to its will, to create entirely new technol-

ogies. These technologies powered the rise of AI, yielding new and transformative yet sometimes fragile artifacts that we barely understand. In turn, AI is turning back to help its human creators understand the recesses of their own minds.

The advances of AI raise profound philosophical questions about the very nature of human meaning and understanding. Indeed, humans are often very uneasy about phenomena they do not understand, an instinct likely built into us by evolution and essential for our survival. Our discomfort with not understanding AI is therefore natural, given its immense impact. But what would such understanding even look like?

Physics again offers guidance. Richard Feynman suggested that we understand a physical system if we can say something about the solutions to its underlying equations of motion *without* actually solving them. For example, if the *only* way we could predict fluid flows was by numerically solving the Navier-Stokes equations, then we would be loath to claim that we understand fluid flow in any meaningful human sense. Instead, we have exact solutions in special cases, as well as a hierarchy of increasingly approximate but more insightful solution methods and deep intuitions gleaned from experience. We can say roughly how a fluid will flow in many scenarios even if we don't know what the Navier-Stokes equations are. Similarly, to understand complex AI systems in a human sense, we could develop analogous hierarchies of approximations, known in physics as coarse-graining. Each approximation trades off simplicity against accuracy in explaining aspects of model behavior. Understanding will lie in being able to traverse this hierarchy, to understand which aspects of a distributed nonlinear circuit are important for which aspects of behavior. Such understanding could lead to better and more robust networks through rational design.

There is also a natural duality to the notion of understanding: How do humans understand AI? And how can AI help humans understand themselves and the world? Regarding the latter, there is palpable excitement about AI providing new explorations into nature, especially for the purposes of verifiable applications, like proving theorems, discovering drugs, and designing novel materials. Here the path is clear and can be thought of as a fortuitous interplay between learning and search. Learning adjusts a parametric function to fit data, providing novel predictions. We have spent the last decade scaling up learning. Search involves traversing a large design space to find good solutions. Historically, brute force search has been intractable. What has changed is that learning allows the creation of good guessers that pick good search directions. For example, in theorem-proving, an LLM can guess the next step of a proof, and a formal proof assistant can verify whether this step is correct. If it is, the LLM can continue guessing; if not, the LLM can backtrack or restart. In this fashion, the guesser and verifier can work together to successfully find correct mathematical proofs even in astronomically large search trees containing many incorrect options.

Even more tantalizing is the possibility of recursive self-improvement. Here, as new proofs are found, the LLM can learn to become a better guesser by adjusting its weights in response to each new proof. And by becoming a better guesser, it can find new proofs more rapidly. Moreover, the learning and search processes can be improved via codesign.[31] This synergistic interaction between learning and search could lead to dramatic acceleration in the discovery of novel engineering solutions in any scenario consisting of a high-dimensional design space, including, for example, drugs and materials, and even the codesign of computation and physics, as proposed above. In principle, recursive self-improvement is possible. After all, the evolutionary journey from autocatalytic reaction sets in Earth's primordial soup to the development of human intelligence may well be the grandest example of recursive self-improvement on our planet.

But if the end goal of this improvement process is the engineering application, human understanding along this journey is not strictly necessary (even if immensely helpful). We just need to verify the design specs (*Is the proof true? Does the drug or material work as desired?*). However, in the pursuit of fundamental science, the *raison d'être* for scientists is deep human conceptual insight into the nature of ourselves, our universe, and our role within it. In essence, *the joy of insight*. Can AI deliver such joy? For example, we would love to intuit how consciousness emerges from neurons, how life emerges from myriad chemical reactions, and how new states of matter emerge from different patterns of quantum entanglement. How far can the intertwined processes of learning and search powering today's AI go toward delivering to us this uniquely human understanding and intuition? What role, if any, will deep human understanding even play in an age when combined learning and search may solve many major engineering and societal problems?

My sense is humans have a strong evolutionary drive to understand and create, which will influence AI for science to evolve in a fundamentally human-centered way: humans will develop a powerful scientific understanding of AI, and AI, in cooperation with humans, will deliver profound scientific insight into ourselves and the world. The exact details of how this will unfold is primarily up to human ingenuity and persistence, not AI. Human-AI science collaboration will be one of the great intellectual adventures of this century. It will be the next chapter of the wondrous trajectory taken on our planet, and it will include AI turning its gaze back on us to help us understand the very nature of our own mental reality.

ABOUT THE AUTHOR

Surya Ganguli is Associate Professor of Applied Physics, Senior Fellow at the Stanford Institute for Human-Centered AI, and Associate Professor, by courtesy, of Neurobiology and of Electrical Engineering at Stanford University. He has recently published in *Neuron*, *Nature*, and *Physical Review X*.

ENDNOTES

[1] Philip W. Anderson, "More Is Different," *Science* 177 (4047) (1972): 393–396.

[2] John J. Hopfield, "Neural Networks and Physical Systems with Emergent Collective Computational Abilities," *Proceedings of the National Academy of Sciences* 79 (8) (1982): 2554.

[3] David H. Ackley, Geoffrey E. Hinton, and Terrence J. Sejnowski, "A Learning Algorithm for Boltzmann Machines," *Cognitive Science* 9 (1) (1985); and Geoffrey E. Hinton and Ruslan R. Salakhutdinov, "Reducing the Dimensionality of Data with Neural Networks," *Science* 313 (5786) (2006): 504–507.

[4] Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, et al., "Hopfield Networks Is All You Need," arXiv (2020), https://doi.org/10.48550/arXiv.2008.02217.

[5] Marc Mézard, Giorgio Parisi, and Miguel Ángel Virasoro, *Spin Glass Theory and Beyond: An Introduction to the Replica Theory and Its Applications* (World Scientific Singapore, 1987).

[6] Yann N. Dauphin, Razvan Pascanu, Caglar Gulcehre, et al., "Identifying and Attacking the Saddle Point Problem in High-Dimensional Non-Convex Optimization," in *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, ed. Z. Ghahramani, M. Welling, C. Cortes, et al. (MIT Press, 2014), 2933–2941.

[7] Yasaman Bahri, Jonathan Kadmon, Jeffrey Pennington, et al., "Statistical Mechanics of Deep Learning," *Annual Review of Condensed Matter Physics* 11 (2020), https://doi.org/10.1146/annurev-conmatphys-031119-050745.

[8] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli, "Deep Unsupervised Learning Using Nonequilibrium Thermodynamics," *Proceedings of the 32nd International Conference on Machine Learning* 37 (MLR Press, 2015).

[9] Mason Kamb and Surya Ganguli, "An Analytic Theory of Creativity in Convolutional Diffusion Models," *Proceedings of the 42nd International Conference on Machine Learning*, ed. Aarti Singh, Maryam Fazel, Daniel Hsu, et al. (MLR Press, 2025).

[10] Jared Kaplan, Sam McCandlish, Tom Henighan, et al., "Scaling Laws for Neural Language Models," arXiv (2020), https://doi.org/10.48550/arXiv.2001.08361.

[11] Francesco Cagnetta, Allan Raventós, Surya Ganguli, and Matthieu Wyart, "Deriving Neural Scaling Laws from the Statistics of Natural Language," arXiv (2026), https://doi.org/10.48550/arXiv.2602.07488.

[12] Ben Sorscher, Robert Geirhos, Shashank Shekhar, et al., "Beyond Neural Scaling Laws: Beating Power Law Scaling Via Data Pruning," in *Advances in Neural Information Processing Systems 35 (NeurIPS 22)* (Curran Associates, Inc., 2023).

[13] Agrim Gupta, Silvio Savarese, Surya Ganguli, and Li Fei-Fei, "Embodied Intelligence via Learning and Evolution," *Nature Communications* 12 (1) (2021): 5721.

[14] Luca Peliti and Simone Pigolotti, *Stochastic Thermodynamics: An Introduction* (Princeton University Press, 2021).

[15] Sarah E. Harvey, Subhaneil Lahiri, and Surya Ganguli, "Universal Energy-Accuracy Tradeoffs in Nonequilibrium Cellular Sensing," *Physical Review E* 108 (1–1) (2023): 014403.

[16] Kevin Mann, Stephane Deny, Surya Ganguli, and Thomas R. Clandinin, "Coupling of Activity, Metabolism and Behaviour across the *Drosophila* Brain," *Nature* 593 (7858) (2021): 244–248.

[17] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and Harnessing Adversarial Examples," arXiv (2014), https://doi.org/10.48550/arXiv.1412.6572.

[18] Atsushi Yamamura and Surya Ganguli, "Fooling LLM Graders into Giving Better Grades through Neural Activity Guided Adversarial Prompting," arXiv (2024), https://doi.org/10.48550/arXiv.2412.15275.

[19] Alessandro Salvatore, Stanislav Fort, and Surya Ganguli, "Solving Adversarial Examples Requires Solving Exponential Misalignment," arXiv (2026), https://doi.org/10.48550/arXiv.2603.03507.

[20] Sadegh Ebrahimi, Jérôme Lecoq, Oleg Rumyantsev, et al., "Emergent Reliability in Sensory Cortical Coding and Inter-Area Communication," *Nature* 605 (7911) (2022): 713–721.

[21] Brendan P. Marsh, Yudan Guo, Ronen M. Kroeze, et al., "Enhancing Associative Memory Recall and Storage Capacity Using Confocal Cavity QED," *Physical Review X* 11 (2021), https://doi.org/10.1103/physrevx.11.021048; and Brendan P. Marsh, David Atri Schuller, Yunpeng Ji, et al., "High-Capacity Associative Memory in a Quantum-Optical Spin Glass," arXiv (2025), https://doi.org/10.48550/arXiv.2509.12202.

[22] Brendan P. Marsh, Ronen M. Kroeze, Surya Ganguli, et al., "Entanglement and Replica Symmetry Breaking in a Driven-Dissipative Quantum Spin Glass," *Physical Review X* 14 (2024): 011026.

[23] Yoshihisa Yamamoto, Timothee Leleu, Surya Ganguli, and Hideo Mabuchi, "Coherent Ising Machines–Quantum Optics and Neural Network Perspectives," *Applied Physics Letters* 117 (16) (2020); Atsushi Yamamura, Hideo Mabuchi, and Surya Ganguli, "Geometric Landscape Annealing as an Optimization Principle Underlying the Coherent Ising Machine," *Physical Review X* 14 (2024); and Federico Ghimenti, Adithya Sriram, Atsushi Yamamura, Hideo Mabuchi, and Surya Ganguli, "The Geometry and Dynamics of Annealed Optimization in the Coherent Ising Machine with Hidden and Planted Solutions," arXiv (2025), https://doi.org/10.48550/arXiv.2510.21109.

[24] Niru Maheswaranathan, Lane T. McIntosh, Hidenori Tanaka, et al., "Interpreting the Retinal Neural Code for Natural Scenes: From Computations to Neurons," *Neuron* 111 (17) (2023): 2742–2755.

[25] Konstantin F. Willeke, Kelli Restivo, Katrin Franke, et al., "Deep Learning-Driven Characterization of Single Cell Tuning in Primate Visual Area V4 Unveils Topological Organization," bioRxiv (2023), https://doi.org/10.1101/2023.05.12.540591.

[26] James H. Marshel, Yoon Seok Kim, Timothy A. Machad, et al., "Cortical Layer–Specific Critical Dynamics Triggering Perception," *Science* 365 (6453) (2019), https://doi.org/10.1126/science.aaw5202.

[27] Jacob M. Hull, Surya Ganguli, and John R. Huguenard, "Interpretable Machine Learning Identifies an Emergent Absence Seizure Mechanism," bioRxiv (2025), https://doi.org/10.1101/2025.09.23.678032.

[28] Ebrahimi, Lecoq, Rumyantsev, et al., "Emergent Reliability in Sensory Cortical Coding and Inter-Area Communication."

[29] Antonio R. Damasio, *The Feeling of What Happens* (Houghton Mifflin Company, 1999).

[30] Ibid.

[31] Feng Chen, Allan Raventós, Nan Cheng, et al., "Rethinking Fine-Tuning When Scaling Test-Time Compute: Limiting Confidence Improves Mathematical Reasoning," arXiv (2025), https://arxiv.org/abs/2502.07154, presented at NeurIPS 2025: The Thirty-Ninth Annual Conference on Neural Information Processing Systems, San Diego, California, December 5, 2025.