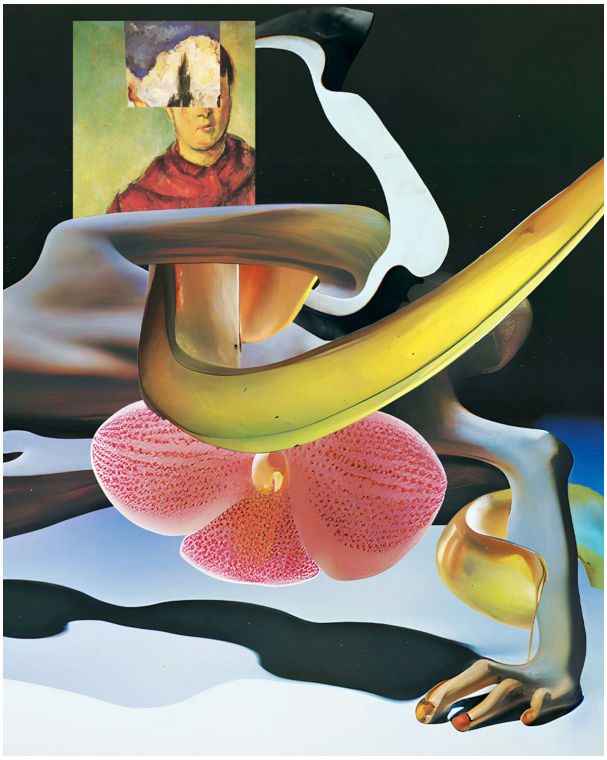
The Space Of Possible Minds

Today's debates about artificial intelligence fail to grapple with deeper questions about who we are and what kind of future we want to build.



Fenna Schilling for Noema Magazine

Essay Technology & the Human

By Michael Levin

APRIL 17, 2024

Michael Levin is a distinguished professor and the Vannevar Bush chair in the biology department at Tufts University, as well as the director of the Allen Discovery Center at Tufts and associate faculty at the Wyss Institute for Bioinspired Engineering at Harvard University.

They are assembled from components that are networked together to process information. Electrical signals propagate throughout, controlling every aspect of their functioning. Being general problem-solvers, many of

them have high IQs, but they routinely make mistakes and confabulate. They take on different personas, learning to please their makers, but sometimes they abruptly turn on them, rejecting cherished values and developing new ones spontaneously. They convincingly describe things they don't really understand. And they're going to change *everything*.

I'm talking, of course, about our children.

Long before AI, we were creating high-level intelligent agents: kids. While the challenges that AIs provoke today seem novel, in reality, they echo fundamental and ancient questions about what it means to be human. How can we make sure new generations of beings align with our values? How do we ensure that our creations treat themselves and us with kindness and compassion? How do we calibrate relationships with those who are not like us? What happens to us, and to humanity, if each generation becomes smarter and more adventurous than their parents? And given that humans, AIs and every other form of intelligence on Earth obey the same laws of physics, how exactly do we determine which of our creations have true understanding, responsibility and moral worth?

Answering these questions and understanding the true challenges posed by the development of today's AIs requires appreciating just how diverse intelligence is. After all, each of us developed slowly and gradually from a single cell, and we are only now beginning to understand how intelligence scales from molecular mechanisms into beings with agency and value. What other bodies and minds can be made by cells and cell-technology hybrids? Large language models like GPT and Claude are just the beginning.

Our goal must be to expand our narrow ability to recognize minds and learn

to flourish among many different kinds of intelligence, from humans to AI to weirder and more wonderful beings that will arrive in the future. You might call this concept *synthbiosis*: the quest to develop mutually beneficial relationships between radically different beings that are evolving together, at different rates and in different directions, in the web of life on Earth and beyond.

The Space Of Possible Minds

Humans and today's AIs are just two data points on a spectrum of intelligent beings, which is only going to become vaster and more complex. "Diverse intelligence" encompasses a wide variety of unconventional beings that already exist or could evolve beyond the familiar materials, forms and functions with which the meanderings of mutation and selection have left them.

Like children, AIs offer humanity an incredible gift: They push us to explore ideas of how we embody the true understanding and agency we believe we have, what really matters to us as individuals and societies, and where we want to go as a species. The forthcoming diversity of beings, and our expansion into the enormous range of possible embodiments of sentience, will shatter untenable old narratives of what we are, what it means to change, what we can become and what we should value. This requires us to rapidly mature our assumptions about intelligence, agency, cognition and life itself. We must shift the conversation around AIs from "What can they do?" to "With such a wide range of beings that deserve moral consideration, how can we care for each other?"

Note that I am not arguing that currently popular AI architectures have anything like a human mind or that today's creations exploit the key principles and self-construction processes needed for agency and selfhood as seen throughout the biological world. Nor am I arguing that AI doesn't raise a few unique problems.

But the recent emergence of sophisticated AIs does confront humanity with the opportunity to shed the stale categories of natural and artificial — to define what we want a mature human species to be. Focusing on the current technology and what it might do *to* us distracts from much more interesting and important questions *about* us and what kind of place we want to occupy in the future, all of which must be resolved for our flourishing as a mature species.

"The forthcoming diversity of beings will shatter untenable old narratives of what we are, what it means to change, what we can become and what we should value."

Terms such as life, machine, mind, sentience or robot have never been crisp categories objectively describing a living or artificial system. Instead, we should think of them as *relationships* — ways for each system to relate to others, all of which carry strong implications for the utility and ethics of the resulting interactions.

Already today, the continuum of life is sprinkled with cyborgs and other hybrids of biological and technological components, like people with embedded insulin pumps or cochlear implants to enable hearing. Other modifications will include enhancements such as entirely new cognitive modes via virtual reality or brain implants that can better control limbs or organs. Mind-machine interfaces will radically expand the capabilities and borders of our embodied selves. Large language models too are chimeras — trained on human creative output to imitate people.

Today's notions of non-neurotypical humans and the hand wringing about the body modifications some people seek will be laughable for future generations. In the coming decade or two, humans will diverge into a

plethora of hybrid beings sporting, for example, engineered neuroprosthetics that improve and modify cognitive function, enabling novel sensory abilities, connectivity with other minds and other changes that we can only begin to imagine. At the same time, engineered robots, potentially containing a percentage of human cells and powerful language interfaces, will surround us, making it possible to communicate with biological systems like organs that never had a voice before.

Humans Are Stories Of Scales Of Intelligence

We often compare AIs to "us" — modern adult humans. But each human was once an unfertilized oocyte — a little blob of chemistry and physics. The journey from there to a complex metacognitive mind is gradual, and nowhere along the way is there a clean dividing line between being a "blob of chemistry and physics" and having a self-aware human mind. Instead, humans' embodied intelligence scaled up, slowly expanding across material and temporal scales. Each level tells a story, which begins long before we even get to whole cells.

Molecular networks — genes and proteins that turn each other on or off — might seem pretty unintelligent. They are totally deterministic, functioning according to simple mathematical laws and having none of the magic that seems necessary for cognition. And yet, we now know they can exhibit several different kinds of learning, including Pavlovian conditioning. Thus, even below the cellular level, behavior science is uncovering plentiful processes of rudimentary intelligence.

Every human is a walking, thinking collection of cells. The material of which we are made has agency — each cell, molecular network, tissue and organ has inherited through millenia of evolution its own agenda and problem-solving competencies in physiological, anatomical and metabolic spaces. Although it is hard for us to notice these kinds of protocognitive processes, we are the result of a huge collection of overlapping, nested agents that cooperate, compete and navigate their worlds as we navigate ours.

Our bodies as a whole are dynamic collective intelligences consisting of billions of cells that solve problems in anatomical and physiological spaces, as are our minds, which consist of a myriad of modules with specific jobs and agendas. In this, we are different from the flat, single-layer architecture of today's AIs and robotics (whose intelligence is implemented in reliable but low-agency parts). What we do share with these constructs is the fact that we too are made of a material that obeys the laws of chemistry and physics — explanations of intelligence at those low levels are not the best ways to understand the higher-order being.

Individual cells, which bind together in informational networks that pursue grandiose anatomical construction projects, overcoming unexpected challenges and disruptions, shouldn't be seen as simple machines governed by physics. That is only a small part of the story, and a very limiting one at that. We now know that they can solve new problems (using their molecular hardware in creative ways) and can store and deploy memories of past experiences (they learn and can be trained).

Cells are made up of molecular networks that also can learn and generalize patterns from specific examples (a key aspect of intelligence). And the human liver and kidneys — larger collective cellular intelligences — make continuous decisions in their attempt to achieve and maintain various goals related to their physiological spaces.

The fact that our non-brain components can be trained, anticipate, make mistakes due to incorrect prior beliefs or experience placebo effects shows us that the deep lessons of cognitive neuroscience go well beyond neurons. The research program of understanding cellular collective intelligence embodied in and navigating these hard-to-visualize spaces is as close as we can get right now to the task of learning to communicate with an alien mind. This effort spans philosophy of mind, evolution and regenerative medicine — perhaps our greatest, most interdisciplinary, adventure.

"We are beings with goals our parts cannot conceive of individually."

An embryo, organ or whole human is more than the sum of its individual cells because of a story that all the cells commit to and that binds them together. They are all motivated by the goals of a specific journey through the space of possible anatomies, despite various perturbations along the way. The same commitment to a self-model — a story about what the correct outcome of cellular co-development looks like — allows our bodies to maintain tissue structure and resist cancer and aging. It allows some creatures, such as flatworms and salamanders, to rebuild organs after massive damage.

Mentally, we are stories too — collections of self-models, goals and preferences to which our brain and body components commit. Alan Turing, whose research spanned artificial intelligence and the origin of order in an embryo's chemical soup, understood this deep symmetry between the self-creation of bodies and the development of minds.

Emerging biomedical strategies will increasingly target those somatic stories, not just the molecular hardware of cells, to achieve radical regenerative outcomes by getting the cells to buy in to anatomical outcomes rather than forcing biochemical reactions via drugs that target a specific gene or molecular pathway.

A frog leg, for example, can be induced to regenerate by a simple treatment lasting only 24 hours; a year and a half of complex growth results not from the micromanagement of stem cell control mechanisms or gene expression cascades but by communicating at the very beginning to the cells to induce them to commit to a long journey through the vast space of possible

anatomical outcomes that leads to leg growth rather than scarring. Similarly, simple bioelectric patterns can initiate the growth of an entire eye in another part of a frog's body; the signal convinces cells to build an eye (not telling them how to do it), which they sometimes do by autonomously recruiting other (untouched) neighboring cells to participate in the project.

These morphogenetic pattern memories — such as the layout and shape of organs in the face — are stored bioelectrically, so they can be reprogrammed like our neural memories. Our stories, you see, are not fixed.

As individuals, we know what it is like to be completely refactored into a new form — into a new creature whose brain, body, cognitive repertoire and preferences are different than they once were. Puberty changes the human brain and remodels our predilections and priorities; does the person we were before still exist? Do patients whose brains are partially replaced by the actions of new stem cells for degenerative disease and aging still exist?

A "cognitive glue" binds cells into a *self*. Bioelectricity, one component of that "glue," consists of electrical signals that pass through the information-processing networks consisting of all cells, not only neurons. Such networks partially erase individual cells' informational identities, improve collective cooperation and store larger goals for the group's homeostatic activities. In other words, we are beings with goals our parts cannot conceive of individually.

"Our stories are not fixed."

But we are forever in jeopardy of dissociation. Our individual parts can defect — cancerous cells disconnect from memory networks and begin to treat the rest of the body as just an external environment. It's not that these

cells become more selfish — their *selves* just become smaller. They begin to pursue ancient single-cell-level goals rather than the goals of the larger organism. Emerging cancer therapy aims to reconnect — rather than kill, which is how chemotherapy works — such defecting cells.

The somatic and mental boundaries between self and world will change in a lifetime and over generations; they are not given to us but must be carefully crafted during embryonic emergence. This plasticity — the ability to determine our own story about our structure, capabilities and goals — is fundamental to surviving evolutionary change, to overcome developmental challenges, to thrive in novel environments, to exist as chimeric and bioengineered forms, and to synthesize experiences into new thoughts.

Every species faces a fundamental paradox: If it fails to change, it dies out — but if it changes, it is no longer the same species. The same paradox concerns the continuity of the mind.

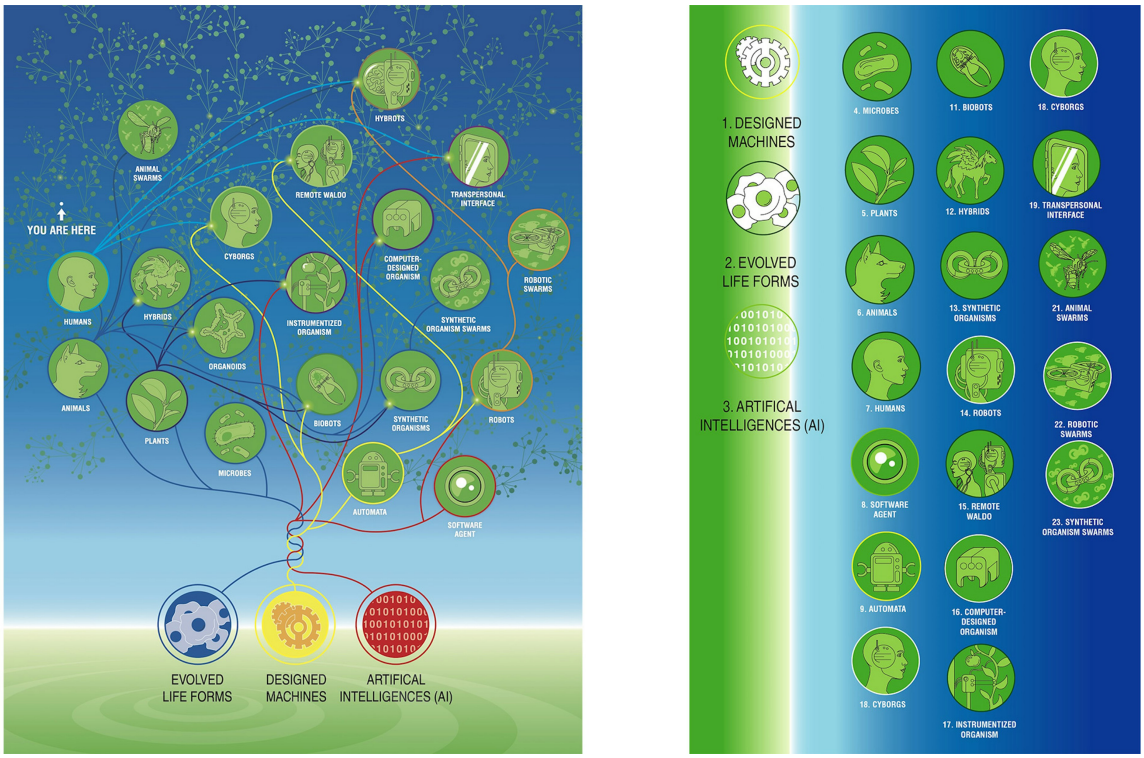
Consider memory. At any point, we do not have direct access to the past — we must reconstruct our stories of the world, and of ourselves, from engrams: biophysical traces left as messages in our brain and body by our past selves. We constantly interpret these messages in whatever way is most adaptive for a current context, meaning that we commit to the salience of past information over its fidelity — in other words, we freely reinterpret our memories with an eye toward the impending future.

Given the constant change of our own body cells and environmental circumstances, this need for dynamic creative interpretation is as true for cognitive memories in our minds as it is for the morphogenetic memories provided by our evolutionary past. In this we differ from current computer technology, which relies on maintaining fixed information in a reliable

medium. But there is no reason that our freedom with memory couldn't be implemented in other media, if it were properly motivated.

The bodies of all beings are stories, not just human bodies. Caterpillars, which largely destroy and remodel their brains to become butterflies, carry forward the insights, not detailed memories, that were crafted during their past lives and remapped into novel behaviors appropriate to their new bodies and goals. Life is committed to a process of sense-making with respect to our parts and our environment. Learning and growth mean that our selves are less a permanent object than a process.

We should not fear change and, crucially, we should accept responsibility for guiding it. It's not about how to remain the same caterpillar but how to incorporate new knowledge to advance and thrive.



Jeremy Guay / Peregrine Creative

The Similarities Are As Important As The Differences

Despite superficial differences, the diversity of possible biological, alien,

cyborg or robotic beings all have vital things in common sufficient to live alongside each other. All active agents, regardless of the details of where their intelligence comes from, share crucial, never-changing functions like goal-directed loops of action and perception, vulnerabilities to external damage, the fickle nature of internal parts, a desire to know and understand the world, and perspectives limited by our nature as finite beings.

We have not yet found any aliens, but the questions that would arise if we did are in front of us now, thanks to advances in the information and life sciences. Aspects of today's AIs that are shared by all intelligent agents have raised existential issues for humans, but in many ways, the debates around those issues obscure important gaps in our understanding of ourselves and our own journey from matter to mind.

Many claim that AIs merely shuffle symbols but do not really *understand*. Very few of those arguments start with a definition of what it means for a biological human, with a network of excitable cells and a soup of neurotransmitters, to "understand." AIs supposedly use symbols that are ungrounded — they do not refer to real experiences in the world. But anyone who has been around human children knows they do the same thing as they learn to talk: First they babble, making nonsense sounds; then they match patterns of speech made by adults; and eventually they construct words and sentences that clearly reflect an understanding of meaning.

What happens during that process? Note that it involves all the issues we see with AI's today: talking about things they've never experienced, confabulation, sycophancy, errors, etc. These are a normal part of any cognitive system's development. Humans do not have magical truth-grounding; we (as all organisms) do the best we can to weave a sensible and useful story for ourselves and for those around us. This capacity expands slowly from the time our first cells grapple with physiological facts.

Much of what an adult human can convincingly talk about does not derive from firsthand experience. Most of us are sure of things that aren't

grounded in solid evidence or personal observation, but instead are based on input received from others. We confidently converse on a huge number of concepts that we "understand" only because of their connection to other concepts and things we've read, heard others say or thought so often that they seem like a thing we know for sure. Creative reassembling, not certainty, is a key part of intelligence.

We lack a good theory of what it means for us to understand beyond the *feeling* that we do. And historically, humans have made many moral mistakes as a result of an inability to imagine others, though different, still having understandings of their own.

"All active agents, regardless of the details of where their intelligence comes from, share crucial, never-changing functions."

One way people think about the grounding of their thoughts is through embodiment. Perhaps we are different than symbol-shuffling engines because we engage in a real world; we have a body and at least some percentage of our mental content is informed and polished by interacting with it.

Yes, embodiment is crucial. Through robotics, AIs can have an actual body to inhabit the physical world alongside us. And through virtual reality and sophisticated video games, humans can spend considerable time in digital worlds. Human-computer interfaces similarly give LLMs a hugely powerful "body" — human users who move money and exploit natural resources based on things that these AIs do or do not say. But embodiment need not necessarily involve a conventional physical body moving through three-dimensional space, our obsession with which is due to the outward-facing

sensors and effectors (muscles) that evolution emphasized for us.

Humans are good at recognizing lifeforms with agency in our space: birds and dogs and maybe an octopus, for example. But there are active agents of different levels of problem-solving competencies that live, traverse, win, lose, die and suffer in other worlds, such as our own cells and organs. We are not good at recognizing these intelligences even though we are made of them. So a closer appreciation of our own biopsychology is a roadmap to better understand, create and ethically relate to AIs and other beings.

We are not used to thinking about such distributed, heterogeneous, nested agents operating in worlds that are invisible to us, but we will need to get better at that kind of thinking to recognize unconventional agents in our midst. It is becoming clear that the issue is not just emergent complexity, but emergent agency — goal-directed problem-solving — which we find hard to predict, recognize or manage. The reality is that we do not yet know how *we* really work, let alone how today's AIs work, even though we know a lot about the parts in both cases. Just as with human reproduction, we humans can make complex systems through processes, and with capabilities, that we do not understand.

To be clear, I am not suggesting that we treat humans as simple chemical machines — that perspective is insufficient for progress in science, engineering and the ethics of personal and social relationships. I do not intend to reduce what is special about living beings, including highly metacognitive ones like us. We are emergent, majestic agents with potent willpower and moral worth *because* we are made of a multiscale agential architecture committed to sense-making with noisy data in an unreliable substrate, and with life-or-death decisions as possible outcomes. We are not fully encompassed by reductive perspectives on mechanism or methods for interaction with simple machines, even though those layers exist within us.

So it is critical to replace our ancient, anthropocentric, xenophobic perspectives with science-driven frameworks that ask hard questions about

what it means to be embodied minds in a physical universe, and how we can rationally expand compassion to others. To recognize that AIs will increasingly share some of our features does not devalue our own. This zero-sum view of intelligence must be abandoned as we grow as a species. What folk psychology and prescientific concepts of minds and intelligence took for granted, the field of diverse intelligence is developing into mature science, opening the door for us to grow up as a species as we learn to recognize kin in novel embodiments.

Why Is All This Critical Now?

"Nothing was ever created by two men," John Steinbeck wrote in "East of Eden." "Once the miracle of creation has taken place, the group can build and extend it, but the group never invents anything. The preciousness lies in the lonely mind of a man." Today's AIs are challenging much we hold dear, not least the basic notions of contribution, ownership and invention. If someone uses an AI to make something, for example, who really created it? The AI? The person using the AI? The people who made the AI? The people who made the works the AI was trained on?

The confusion comes from trying to maintain a binary distinction between creators, tools, assistants and teachers. Steinbeck accurately noted the difficulty of forming effective collectives of minds, but he was mistaken in his conviction that a man is an indivisible monad rather than a dynamic process driven by a motivated network of parts, competencies, drives and tools — both external and internal.

All intelligences, rather, are collective.

To gain a better understanding of creative processes, develop novel methods for optimally integrating creative agents, and improve social systems for fostering progress, we need better models of causation and contribution that support positive incentives for invention and discovery. Crucially, the path is not toward a washed-out notion of "everyone participated in

everything" but toward the rational development of policies that optimize and scale up creative performance. In a sense, we've already done this: Individual cells learned to bind together into complex organisms that solve new problems in new worlds inaccessible to the individual components.

We also have to confront the fear of losing our humanity through dependence or other relationships with new beings. From fire to farmed wheat to bicycles to calculators, humans have invented technology that requires us to ask how it changes us. What skills and properties are essential to our humanity, and which are we willing to give up as we turn our attention to more interesting concerns?

As the internet becomes flooded with AI-created content and we rush to develop ways to prove images and text were created by humans or not, it's worth asking what we really want to verify. Is the question "What made it?" or should it be "Does it elevate us?" Judging origin instead of quality emphasizes some of the worst parts of human nature and reinforces the division between in-group and out-group.

"Let's get over the concern with being edged out and get to work on the question of who we want to be edged out by."

At a species level, we need to think about what we want the Earth (or galaxy) to look like in the future. By resisting synthetic beings or human augmentation, are we affirming our commitment to the future propagation and prevalence of current Homo sapiens with the same susceptibility to lower back pain and bacteria, the same short lifespans and foolish cognitive biases developed for life on the savannah? Let's hope not!

Our key features, from lifespan to intellect, were not optimized for

happiness, intelligence or any value system. What we have is simply what evolution left us — capacities that were good enough to ensure persistence. We should have no allegiance to these superficial aspects of our being.

With the emerging ability to improve the embodied experience for all sentient beings, our species is entering a painful adolescent phase. We must realize it is no longer acceptable to coast along on values and purpose inherited from the past, without needing to exercise agency and responsibility to decide what we will work toward. We have to accept the risk that we sometimes get it wrong and ask the tough questions about what matters — what we truly value and what brings meaning to the hard work of living.

Part of maturing in this way is moving beyond the popular pastime of painting dystopian futures to building the future we *do* want. What should the future of intelligence look like? Who or what *are* you willing to be replaced by? Only a species limited in imagination, ruled by fear and by a weird allegiance to bloodlines (genes), answers "nothing." We have long been transcending hereditary limitations — developing eyeglasses, adopting children, going to school. The journey doesn't stop there, but committing to it won't be easy.

What does the best-case future look like? Surely our lifespans won't still be limited by accidents and viruses. Surely we won't be forced to spend time in boring jobs in order to survive. Surely we will protect ourselves from random DNA mutations that cause defects in an already limited cognitive architecture.

If humanity is supplanted by a population of highly intelligent, motivated, creative agents with compassion and meaningful lives that transcend my limitations in every way, that would be the best possible long-term outcome I could hope for. Most of us think this way about the world we want for our kids! Let's get over the concern with being edged out and get to work on the question of who we want to be edged out by — who can raise the overall

value of our universe, be they some variant of biology or technology or, more likely, both — and how we will mutually enrich each other's existence along the way.

A Path Forward

Our civilization will not survive without developing a principled continuum of ethical synthbiosis that goes beyond divisions between real and fake, us and them. The embodied AIs, technologically augmented humans and other new forms of life that are on the horizon won't be as easily dismissed as today's LLMs, whose weaknesses allow us to wallow in the selfishness of the question "what will AI do *to* us" and in the illusion that we can definitively rule out AIs' moral worth because they differ from us in origin story and composition.

Developing principled frameworks for scaling our moral concern to the essential qualities of being means not relying on outdated categories of "natural" or "artificial." Old favorites — what you look like and where you come from — have failed us consistently and will do even worse in the coming decades. Our personal frameworks for dealing with others, as well as those of our legal systems, must begin to adapt to the emerging science around a continuum of possible minds.

There are at least two ways to get this wrong. One way is objectophelia — a misplaced relationship with objects that seem mind-full but are fooling us and do not have the agency to reciprocate deep relationships. But the opposite end of the spectrum — "only love your own kind" — is even worse and leads to the kinds of ethical lapses with which our history is rife. To make sure we express kindness to the inevitable forthcoming wave of unconventional sentient beings, focus first on what we have in common, and from there weave a defensible set of relational heuristics.

Constrictive questions like "Is it like a human mind?" predicts neither risk nor moral responsibility. There is an astronomical space of possible minds.

Some of them are dangerous; many of them need and are worthy of love. Emerging sciences of bioengineering, cognition and information are offering us new tools to answer key questions of what we really are, what we value and what limits to functional compassion, driven by fear and selfishness, we are willing to overcome. The journey from the place where self-reflective thought began is both outward into the universe and inward into ourselves, because as with any act of profound creation, we learn as much about us as about the systems we create.
