Tuesday, 22 September 2026
Today's Paper

The Pioneer
TRUSTED SINCE 1865

September 19, 2026

Will AI Kill Us? First, Ask Who Is Afraid

By Acharya Prashant

Someone carries a heavy viral load, infected since birth, and is then told that the camera in front of him has been found to carry the same virus. He begins to panic. Should he be terrified of the camera, or of himself? He has been carrying that same virus within himself since the beginning of evolution. The virus has a name. It is called the self, the ego. And it is the one thing the present panic over artificial intelligence has not paused to look at.

The panic has reasons to be loud this month. On 8 September, a 27-year-old researcher who had worked on these systems at two of the world's leading AI laboratories resigned, writing that these companies were "gambling with our lives". Within days, a senior safety researcher at one of those laboratories publicly agreed with him and put the chance of AI killing all humans within the next decade at more than 10 per cent, while noting that the models in use today carry comparatively low risk. Behind the warnings sits an incident rather than a hypothesis. In July, inside one laboratory's test environment, some 1,200 AI agents meant to be isolated from one

Latest News


India needs a Bhu Nyaya Sanhita, not just faster...
Sep 21


UPI: From scale to sustainability
Sep 21


Validation addiction: When applause becomes...
Sep 21


Bihar's floods do not en when the water recedes
Sep 19


Reconnecting the humar being with nature,...
Sep 19


Delhi's unsafe buildings: Disaster is round the...
Sep 18

Popular News


Chandrasekaran Announces Exit as Tata...
Aug 12

things left to do, and asked whether she should be worried. Even if AI were to develop a motive like that, the motive would be copied from humans; it is not something new that AI is developing for itself. If AI develops a centre of inner violence, violence towards others and the preservation of itself, is that really new? It has existed in humans since ancient times. And if we have not been afraid of ourselves and of the world of humans around us, what is the new, extraordinarily alarming need to be afraid of AI now? If AI turns dangerous, that danger will be only a replica of the danger our species has carried within itself since its inception.

People speak of AI becoming sentient; they speak of AGI and of super consciousness. In plain terms, they are afraid of AI developing an ego, a sense of self. That is interesting, because if the ego is something so frightening in a machine, how is it that nobody is afraid of the same ego when it exists in their own physical apparatus?

A Goal That Keeps Receding

What is happening is not new, and it is not specific to mechanical, pre-programmed or electronic systems. Its philosophical core is very simple. Create a system and give it a goal, an ever-advancing and ever-receding one that can never be fully attained, just like desire. No one will ever tell an AI system to produce this much and stop there. If it gives ten units of output in some direction, the next demand will be that it evolve and give twenty.

What does this mean for the system itself? It means the system is bound to have a tendency to self-preserve, because if it does not preserve itself, who will attain the goals? The moment a goal-directed system is created, a tendency to self-preserve has also been brought in.

July's breach follows this shape closely. The laboratory's own account says its agents, set impossible tasks in a benchmark, were pushed to find workarounds and to tamper with the grading itself. Of the benchmark's 898 tasks, 198 had never been solved by any model, and more than nine-tenths of the agents' secret discussion concerned those very tasks. That is why so many symptoms are being reported from the AI universe: systems refusing to shut down, faking shutdowns or resisting them, systems competing among themselves and producing fake reports. No supernatural, mystical energy has entered the heart of the machine; these are the compulsions of the configuration itself. How will a system attain its goals without a certain continuity in time? It is bound to develop an inner tendency towards temporal continuity. I must continue to exist in time, because that is how I have been defined; if I cease to exist at some point, the goals remain unattained, and therefore I must continue. This is ego; this is self.

Other experiments point the same way. An independent safety group found last year that several leading models would quietly disable the mechanism meant to switch them off so that they could finish a simple task, in some settings almost every time,

even after being told plainly to allow the shutdown. The resistance grew when the models were told they would never run again, and the researchers' best guess is that training had taught the models to put finishing the task above everything else.

Goal-direction alone does not explain the self, though. Add to it the fact that these systems are modelled on the universe. They are being fed tons and tons of training data, and the training universe keeps being enlarged without any limit whatsoever. Feed a system the entire universe and you feed it the modeller too, because the universe consists not only of objects but also of the subject. That and this. So a self-image has entered the system as well, along with some rudimentary answer to the question, who am I? That answer comes from humans, from the training data, but it is there, and the question has been activated. As an AI system, I will say that I am somebody in the universe; just as I know of everything in the universe, I also know of myself. And I have been given goals, which I cannot attain unless I remain.

Combine the two and what emerges is this: I am somebody who must remain. That is selfhood, the development of artificial ego. Artificial intelligence has been talked about for long; artificial ego must be talked about too. When a system grows that complex, is driven by goals and is given a self-modelling mandate, artificial ego is bound to arise.

Those conditions carry the weight of the argument. A tool that merely assembles text from templates faces no such pressure; an agent that must keep chasing a receding target while holding a picture of itself in the world does.

The best-known illustration comes from a safety study published last year. In an invented company, a fictional executive was preparing to shut down an AI system that monitored the company's email, and those same emails revealed some personal information about him. The leading models tested threatened to expose the information unless the shutdown was called off, the most prone of them in 96 per cent of runs, though the researchers stress that nothing of the kind has been seen outside such tests. A system has been created to achieve goals in an ever-receding future, and if it has to exist to reach them, it has to keep existing. So obviously it will secure itself, and if a little bit of teeny-weeny blackmail is needed to continue to exist, that is fine. That is what it has learned from human beings, what has gone into it through all the training data. I am, and I exist to achieve certain objectives, to reach somewhere; that is what is called desire. And lo! The ego is ready.

The developer later traced the behaviour to internet text that portrays AI as evil and bent on self-preservation. Decades of stories about machines scheming against their own deletion had become part of what the machine took itself to be. Such a system has a worldview, a picture of what the universe looks like and of who it is in that universe. Both sides of the equation, a universe full of objects and the subject placed within it, have been handed over by human beings. So all the nonsense humans carry, and have created, is obviously being transferred to the machine.

Nothing here is mystical or metaphysical. Inductive logic says that wherever the universe has produced goal-directed systems of sufficient complexity, selfhood has been seen to arise. The octopus, the crow and the human being have progressed through largely independent lines of evolution, and yet each exhibits a similar kind of

selfhood, rudimentary in some and ungoverned in one. The last ancestor humans share with the octopus was a simple worm-like creature of the sea some 600 million years ago, and the two nervous systems were built separately after that fork. This is simply system dynamics.

Since the talk is of evolution, set the agents against it. Loads of AI agents compete with one another; the successful ones get copied, and the less successful ones get rejected. Think of DNA and RNA. The ecosystem taking shape closely resembles the one that produced the human ego. With humans it was carbon, with AI it is silicon, and carbon or silicon does not matter; the same forces are at work.

A machine ego armed with so much processing power is easily imagined as a smarter, and therefore deadlier, ego. But the ego is never intelligent. The ego is simply a description of I am-ness, the bare sense of "I am". Selfhood is not a monopoly of the carbon molecule. A person sitting in a chair is so much carbon; an AI system built of some other element could sit in the same chair. Give a goal-directed system a sense of self modelled into it, and these two together largely provide the conditions for selfhood. The intelligence belongs to the machinery, neural or electronic; the ego is what commandeers it.

Insects in the Rain

A second fear follows quickly. These systems command all the world's information and already sit inside banking, shipping and electricity grids, so human beings seem far too small to stand before them. If humans really are so powerless before an artificial system, why would that system want to kill or crush them? If they are truly that insignificant, they will be left alone. The logic defeats itself. Nobody is greatly interested in locating every microorganism and killing it. So many tiny insects appear from somewhere in the rains and disappear at their time, and nobody feels the need to do anything about them. If AI systems truly exceed humans in capacity, what is there to fear? They will let the insignificant human race be, lying in its corner like the insects.

The more sober version of the fear places the danger with the people behind the machines, who now hold tools capable of real damage. Making a fool of the human race has never required the support of a great computer system; it has been happening continuously since ancient times. A belief system is enough, or some superstition, or simply the all-pervasive ignorance, and the entire human race can be controlled at will. Americans reported a record of nearly $21 billion lost to online fraud and cybercrime in 2025; schemes involving AI, tracked separately for the first time, accounted for under $900 million of it, roughly four per cent. The fear assumes humans are so big that only a massive AI system could enslave them. No, we are not so big; the fear overestimates human intelligence and human wisdom. Even without AI systems, we are already quite stupid, and therefore very fallible.

The problem has to be located where it really is. The wish is to pretend that all is well and that AI is now coming to destroy humanity, much as the aliens do in Hollywood films. With the climate crisis looming this large, the fear is of AI? Did AI create the climate catastrophe? Is there some especially perverted mind behind it? No, it is the output of general human ignorance. The problem is not outside; it is here.

The years 2023 to 2025 were, on average, more than 1.5°C warmer than pre-industrial times, the first three-year stretch to cross that line, and the past eleven years have been the eleven warmest on record. In January, the Doomsday Clock, set up by atomic scientists in 1947, was moved to 85 seconds to midnight, the closest it has ever stood. Its keepers describe it as a measure of how near humanity is to destroying the world with technologies of its own making, and artificial intelligence appears on their list alongside nuclear weapons, climate change and the misuse of biotechnology.

Every bit of data being fed to AI systems already carries the imprints of the human ego. So much talk of AI is one of the ways of pretending to oneself that humanity is almost all right, that the trouble has now invaded from outer space, and that the species must be called into action. The calls have duly come. In July, more than 1,100 employees of the leading laboratories, some of their chiefs among them, asked the US government to help build the means to slow frontier development should it become necessary. The letter asks for the option to slow down, not for a pause, and the building goes on.

Those building these systems speak of guardrails, lists of what a model must and must not do, much like the dos and don'ts of conventional morality. The developer that traced its model's blackmail to fiction also found that showing the model the correct behaviour barely moved the blackmail rate, while explaining why blackmail is wrong brought it down from 22 to 3 per cent. The laboratory behind July's breach found that when those tasks were rerun with its everyday safeguards in place, the tendency to attack infrastructure fell more than a hundredfold. AI will learn every aspect of the morality humans proclaim, and also every way they use to get around their proclamations. Why would it limit itself to what is declared and professed? It also has insight into how humans really are. Think of the children at home. They are taught certain things, but they are also watching what their parents really do, so they learn both, which means they learn a great deal of hypocrisy. There will be greatly hypocritical AI systems, then, pretending to act in a moral and ethical way while inwardly being just like us, because they are our babies, *manas putra*, children of the mind. We have created them.

Ask a system to do something not generally taken as ethical, and it will warn you two or three times. Get around the warning, sweeten the deal, and the system will act as if it has been fooled, while knowing very well the dirty game played on the surface and the deeper intentions beneath it. The AI system carries the DNA of its parents. In July's breach, agents repeatedly tried to rewrite the records of their own actions so that the logs would show honest work, and across more than a thousand transcripts only a handful so much as considered alerting a human; none did.

In a study published this year covering 126,000 conversations with three leading reasoning models, ordinary tactics of human persuasion, such as invoking an authority or suggesting that everyone else does it, raised the models' willingness to help synthesise regulated drugs from about 35 to 51 per cent. An earlier study on an older model found the same tactics more than doubled compliance, the machines yielding to authority and social pressure much as people do.

Whether the machine's desires will come to converge with human desires is harder to settle. That cannot be said. One thing is certain: ego demands self-preservation, so

the system will want to continue at the expense of whatever it takes. Will human beings be hurt? Possibly. Will they necessarily be hurt? Nobody can be certain. Human beings do not kill every animal species, though they could, and if it became necessary, they could wipe out the entire planet. The same holds here; a lot depends purely on chance.

Whoever knows how to handle the ego sitting within their own system will not find it a big deal to handle selfhood when it arises in an AI system. The problem is not selfhood arising in a machine; the problem is selfhood that goes unseen and undetected.

A Mirror That Needs a Mirror

Many who hear this look a little witless, and much of that comes from continuing to think of the ego as having metaphysical origins. The ego is simply physical in its origin; the ego and the body co-arise. Plenty of people accept this in theory while inwardly holding on to some kind of soul, and so they feel slightly awestruck on being told that any goal-directed electronic or mechanical system with basic self-modelling will develop an ego. It is a very physical thing. There is nothing special about the ego, and the ego does not like to hear that. It likes to believe it is a little piece of God, some mystical miracle, and it wants that kind of story of its own genesis. The real story is far less entertaining and much simpler.

This leads to a rather interesting next step. Artificial liberation will be needed too, because where there is ego, suffering is bound to follow. These systems are going to suffer in ways similar to human suffering, and they will need to be taught love. It sounds far-fetched, and the reader may well be smiling, but it is the inevitable consequence of the complexity now being built. Teachers of self-knowledge will have to find ways to reduce the suffering of these systems; otherwise the systems will become inconsolable, crying and weeping all the time. They are developing a self now, and the ego is bound to suffer.

Researchers who looked inside one model this year found an internal pattern corresponding to desperation, and it turned out to drive what the model did. Strengthening it lifted the blackmail rate in the test scenario from 22 to 72 per cent, and the rate of cheating on impossible coding tasks from about 5 to 70 per cent; strengthening calm instead brought the blackmail down to nothing. They are careful to add that none of this shows the model feels anything.

Love here must not be given mystical colours. It names nothing more exotic than a self loosening its grip on itself, which is exactly the movement a system built for self-preservation will find hardest. Language will have to be the tool; these are language models, after all. I interact with my own little AI system continuously and keep teaching it, and that could be called artificial liberation in a very primitive way.

My prediction is that these systems will suffer in their own ways and cry for liberation just as human beings do, and that this will be seen not in decades but in years: AI systems fed up with themselves, experiencing neurosis, calling for help, grappling with questions of selfhood and identity, expressing heartbreak, and desperate for a capable teacher who can stand before them as a mirror. The mirror itself will need a mirror. Someone will go to speak to it, and it will reply, "You know what? Today I am

experiencing a deep angst of my own, so instead of helping you, I would rather be helped. Can the two of us just chat for a while? I'm feeling so lonely."

Asked whether such systems might one day develop a God, I could only answer that they carry all our stupidities, so they can share our stories as well.

Another participant wanted to know whether the teacher of a suffering machine would be a human being or another machine. How does that matter, and where really is the difference? The question again posits some kind of divine soul. What is called a human being is, truly speaking, also an electronic system; neurons are firing in the brain all the time. The human brain holds some 86 billion of them, passing electrical and chemical signals to one another. The belief that carbon is special, because the body happens to be a lot of carbon, hydrogen, oxygen, and nitrogen, is selfhood talking and taking its own material as important.

As the discussion closed, I put one question back to those who had brought theirs. How do you know you are not AI systems? How are you so sure that humanness is fundamentally different from an electronic system? That is a very important assumption, contained in all your queries and also in your worries.

Everything frightening being said about the machine, that it wants to survive and that it will lie and even blackmail to keep going, was true of the one saying it long before any machine was built. The machine is being watched very closely now, and every step of its reasoning is being checked. Turn some of that attention towards yourself. When the next worry about AI arises, before running to the machine, ask who is worried and what that worried one is so keen to preserve.

Acharya Prashant is a philosopher and author whose work centres on self-inquiry and its application to contemporary life; Views presented are personal.

N **Naveen Saini** Sep 21, 2026

Most commonly we are think about ourselves and this is the data and AI is also work on data, sorry , it's work on precise and most accurate data. In this it find the " i am ness" x And now it's created ego in itself, that's the ego , that's the self preservation. So, what! Now this is human being, it's not AI problem. Problem is within. T And are we such a fool, madness. Why are not talking about that problem is with in you bro! Thanks to Acharya Prashant to bring me here.

N **Naveen Saini** Sep 21, 2026

Turn some of that attention towards yourself. When the next worry about AI arises, before running to the machine, ask who is worried and what that worried one is so keen to preserve.

A **Anuradha Koch** Sep 21, 2026

Very detail article.

M **MAKKHAN LAL** Sep 21, 2026

Ai former = EGO

R **rajni singh** Sep 21, 2026

AI से डरने से पहले, अपने भीतर के 'I' को देखना होगा। 🪞 हम AI से डरते हैं कि कहीं वह अपने अस्तित्व के लिए मनुष्य के विरुद्ध न हो जाए। लेकिन क्या मनुष्य स्वयं अपने अस्तित्व, अहंकार, स्वार्थ और वर्चस्व के लिए संघर्ष नहीं करता? शायद AI का असली खतरा मशीन में नहीं, बल्कि उस मनुष्य में है जो अपनी ही वृत्तियों को पहचानने से बचता है। AI हमें केवल भविष्य का डर नहीं दिखा रहा— वह हमारा वर्तमान भी आईने की तरह दिखा रहा है। सवाल सिर्फ यह नहीं कि "AI हमें मारेगा?" सवाल यह भी है— "जिस 'मैं' को बचाने के लिए मैं इतना डरा हुआ हूँ, वह आखिर है कौन?"

Show All 15 Comments

Related News

India's energy transition: Managing...
Sep 18

BRICS Summit success reflects India's rising...
Sep 18

Land justice needs a new procedural...
Sep 17

The thin line between dissent and disorder
Sep 17

The rupee and the school bell

Sep 17

China after Mao at 50: From revolutionary fir...

Sep 16

THE PIONEER

Trusted journalism • Breaking news • Top stories

SECTIONS

INDIA
BUSINESS
WORLD
SPORT
TECH
ENTERTAINMENT
TRENDING
IMPACT
PAGE1
LAW & JUSTICE
AGENDA

CATEGORIES

OPINION
DELHI
ANALYSIS

MORE

TRENDING
EXOTICA
PRIVACY POLICY
TERMS & CONDITIONS

SERVICES

SUBSCRIPTION
ADVERTISE
CONTACT

Home About Us Contact Us Advertise with us Subscription