THE kyndryl institute

Trend Topic: **Artificial Integrity**

By

Professor Osamu Sakura

Evolutionary biologist; Professor, Department of Social Design, Faculty of Human and Social Sciences, Jissen Women's University

Starting from evolutionary biology, Prof Sakura has studied the history of biology and the relationship between science, technology, and society. After serving as Associate Professor in the Faculty of Business Administration at Yokohama National University (1993–2000), Visiting Professor at the University of Freiburg in Germany (1995–96), and Professor at the Interfaculty Initiative in Information Studies at the University of Tokyo (2000–25), he assumed his current position of Special Visiting Professor at the National Museum of Ethnology and Professor Emeritus at the University of Tokyo and Honorary Research Scientist at RIKEN.

Explore more from The Kyndryl Institute
kyndryl.com/institute

Community sustains ethics and resilience in the age of AI

The growing issue of AI and inequality

What makes ethics difficult today is that it is no longer a standalone subject. It only becomes meaningful when considered together with social, economic, institutional and technological structures. That is to say we need to look at the broader forces shaping AI development.

I have heard that at IT companies on the U.S. West Coast, student interns can earn as much as $10,000 a month. Under those conditions, it is hardly surprising that inequality widens. Compared with manufacturing, which involves physical production processes, the IT industry inherently requires relatively less labor to create value, which means the same level of profit can be generated with far fewer people. The more the information industry expands, the greater its economic impact becomes —but the number of people employed remains far smaller than in manufacturing. That is the structure we now face: a handful of tech companies versus everyone else, with the gap growing ever wider. It may be that society has reached the limits of a system built on the belief that competition, and the inequalities it produces, will keep people motivated.

AI development in particular depends heavily on competition over quantity — especially computing resources and data. As a result, development capacity concentrates in the hands of the small number of companies able to make enormous investments. If only a handful of firms are capable of building these systems, can fairness and equity really be safeguarded?

Can governance create a more ethical AI?

When recombinant DNA technology was invented, scientists recognized its risks even as they were excited by its immense potential. They convened an international conference themselves, bringing together around 140 experts from 28 countries to draft guidelines. This was the Asilomar Conference of 1975. It is remembered as a case in which the autonomous governance of science actually worked: researchers questioned their own social responsibility, even at the cost of constraining their own freedom of research.

A similar conference for AI was held in 2017, resulting in the publication of the Asilomar AI Principles. However, the circumstances were different from those surrounding genetic engineering. In AI, the center of technological development has shifted from universities and public institutions to private companies, and the experts with the most advanced knowledge now tend to be employees of OpenAI or Meta. Under those conditions, profit and speed inevitably take priority over building common rules. Decisions about who can use AI models — and under what conditions — are increasingly being made not through democratic debate or government involvement, but at the discretion of tech companies themselves.

The EU offers another approach: strengthening legal systems and regulations. But it is difficult for states and regulators to keep pace with the speed of technological change. If that is the case, then the key may be for society to apply sustained ethical pressure to tech companies. By pressure, I mean a broad social consensus and a persistently critical public gaze. AI is now embedded in critical infrastructure. If a major accident were to shake society, that pressure could conceivably bring AI development to a halt. If developers inside tech companies feel that risk keenly, their ethical awareness may strengthen as they will want to demonstrate clearly that AI can benefit society. After all, it is not that people working on AI in private companies lack a conscience. Some robotics startups, for example, have already created their own independent ethics committees, and I expect such efforts to increase.

As AI grows more advanced, what matters most is resilience

That said, I think it is an illusion to believe that, no matter how many rules we set, everyone will be able to use AI safely. This is true not only of AI. The very idea that human beings can fully control technology is itself an illusion. With nuclear power, airplanes, and automobiles alike, accidents occur with a certain probability, yet society has learned to live with them. What matters is resilience: anticipating that some accidents will happen and thinking seriously about how we recover when they do.

At this stage, it is important to design AI with accountability in mind. Whenever a new technology is introduced to society, the benefits and possibilities tend to be emphasized above all else. But the brighter expectations shine, the stronger the backlash becomes when unexpected trouble arises and people feel, “This is not what we were promised.” Trust can be lost with surprising speed.

The very idea that human beings can fully control technology is itself an illusion.

That is why it is essential to carefully explain the negative side of technology and its risks, and to build those realities into the way society understands and adopts them.

In fact, there is still no adequate social consensus on what the real risks of AI are, or how much of that risk society is prepared to tolerate. It may even be that no single consensus should be expected. Sensitivity to risk differs from person to person, and such diversity may itself be healthy. In practice, distrust of information systems used in finance and critical infrastructure is often tied to memories of past failures or accidents — in other words, to concrete personal and social experience that creates a sense that "if something happens again, things could go badly wrong." Starting from that variability, society must think about what level of risk it is willing to accept.

At the same time, just as human beings cannot fully manage natural ecosystems, we should not assume that people will be able to control every aspect of AI forever. The emergence of extremely advanced models such as Claude Mythos had been anticipated for more than a decade. Going forward, it will become increasingly normal to use AI to defend against AI-driven attacks. Even in chess, contests between AIs have already moved beyond the reach of human understanding. The internal structure and decision-making processes of AI are already becoming black boxes that humans cannot fully understand or explain. So even if we demand transparency, it will be almost impossible to make everything visible, and even if transparency could be achieved, the complexity and sheer scale of the processes involved may make meaningful human oversight unrealistic. In that sense, one possible future is that AI behavior and interactions among AIs themselves, will largely be left to AI rather than to humans. Just as human beings have domesticated or cultivated only a tiny fraction of the genetic resources in nature, we may leave much of this activity to AI while selectively drawing out the parts that are useful to us. And as with natural disasters, what will matter is having the means in place to restore systems quickly when accidents occur. Leaving AI behaviour and interactions to AI is not necessarily a frightening idea. Ecosystems are not simply something to be feared, but nor can they or should they be fully controlled; they are something we coexist with. That is precisely why the question of resilience becomes so important. The point is not to abandon responsibility, but to recognize the limits of control and think seriously about how we coexist with highly advanced AI systems, including how we prepare for failure and recovery.

The arrival of highly advanced models may be a good moment to reconsider the relationship between AI and human beings. The idea of the human in the loop has been widely discussed for the past four or five years, but I have always felt some discomfort with it. In East Asia, there has long been a sensibility that sees human beings and tools as integrated rather than separate. In Japan, for example, there is a long-standing belief that tools used over many years come to house a spirit — in other words, that they evolve like living beings. Human beings and technology are understood as inseparably entangled, so the notion that a person must be inserted into a system afterward is not one that naturally emerges in Eastern thought. If we think

of AI instead as something with which we coexist within an ecosystem, we may be able to move away from a mindset of control. And if we do, then even if far more powerful models appear, we may no longer need to regard them as enemies from the outset.

Capabilities needed in the age of AI natives

In practice, large language models (LLMs) such as ChatGPT have already become something like companions for many people. They are now firmly embedded in daily life as tools that empower individuals in their work — writing emails and performing secretarial or administrative tasks. But this creates a new problem: how to provide training for truly AI-native generations who have never experienced those simple tasks that AI has now taken over.

Human beings acquire language through interaction with the real world, whereas LLMs learn through entirely different algorithms. For that reason, we cannot simply equate AI with human thought or imagination merely because both use language. Even in situations where the validity of a judgment must be checked, AI makes its judgments on the basis of the information currently available to it. Whether it can autonomously generate genuinely new value or direction for the future is, I suspect, still another matter. At least for now, the final spark — the leap of intuition or imagination — remains a human domain, and I believe it will continue to remain part of the human role for some time.

Yet until now, cultivating experts capable of making those judgments and checks has required long periods of repetitive training. In medical imaging, for example, practitioners spend 10 or 20 years looking at thousands of images and training themselves to detect minute differences. But now that AI can make such determinations with extremely high accuracy, younger medical professionals would naturally doubt why they must keep repeating tasks that AI can already do. In other words, the opportunities to build the foundational skills that have traditionally underpinned expertise are disappearing rapidly. The people who use AI most effectively today are those who underwent that kind of drill-based training in the pre-AI era. But expecting future AI-native generations to go through the same training could easily become counterproductive, or even coercive.

So, what does expertise look like in an age when human beings no longer repeat simple tasks themselves? One key lies in cultivating the ability to frame questions. Educational research is increasingly showing that AI can be effective as a kind of intellectual sparring partner — helping people surface issues they would not identify on their own, deepen their questions, and revise them through repeated reflection. As a university professor, I see this shift firsthand: even in higher education, the focus is moving away from training students to produce answers and toward training them to formulate questions. The traditional model — assigning reports and grading the final submission — is no longer working in the age of generative AI. Now we bring students together in one place and assess

how they approach a problem for themselves: where they begin, how they frame it, and how they choose to tackle it. This change also points directly to the kind of people organizations will increasingly need – something I call creative followers.

Thought leadership that supports followership

Virtually every university now promotes a slogan along the lines of "developing next-generation leaders." But the overwhelming majority of graduates will not in fact become leaders. They will work as followers in the middle layers of organizations. What we should really be cultivating are intelligent, creative followers who have the capacity to move organizations from within. In particular, as individuals are empowered by AI and followers themselves become able to identify problems autonomously using AI, what will be required is followership: the capacity, when necessary, to challenge leaders and elevate the performance of the team as a whole.

If that is the premise, then what organizations need in the age of AI is a coordinating form of leadership — leadership that supports and enables the team. It is not a model in which the CEO supplies the answers each time a problem arises and pulls everyone in a fixed direction. In an age of uncertainty, when values, economic conditions, and international relations are all in flux, a dogmatic mindset is no match for that reality. What matters instead is a moderator-like mode of leadership: sharing a purpose, refining the questions brought by the team through dialogue, and moving forward while adjusting course flexibly when needed.

When that happens, what motivates followers will not be short-term performance metrics such as profitability or efficiency. Leaders will need the ability to articulate a larger purpose or value in a way that people can genuinely accept. In that sense, the importance of humanities — history, philosophy, and the arts — may once again be increasing. That is to say, in an era when AI can guide us toward solutions, what human beings need is the capacity for reflection and contemplation.

This way of thinking is deeply connected to how organizations understand resilience. Resilience is not simply the ability to recover. It is the capacity to adjust direction flexibly in response to changing circumstances. To strengthen resilience, broad knowledge and judgment are essential: not only knowledge of AI, but also a wide and substantial grounding in human history and philosophy, or anything else that enriches human thought and language. It is important for leaders to exercise that kind of thought leadership, but no one person can do all of that alone. This is why the idea of community matters. Some people bring deep knowledge of history, others of philosophy, others of technology; their different forms of knowledge are pooled and made mutually supportive. AI may support integrating those forms of knowledge. If organizations can foster such AI-empowered communities, and if those communities can connect with one another, then they may become a force for improving society as a whole.

In particular, as individuals are empowered by AI and followers themselves become able to identify problems autonomously using AI, what will be required is followership: the capacity, when necessary, to challenge leaders and elevate the performance of the team as a whole.