YIN JIE

Eastern wisdom to inspire global AI governance

Artificial intelligence (AI) is entering a historic juncture as its capabilities begin to surpass those of humans. Harnessing AI's technological dividends while safeguarding the foundations of human survival and development and preserving human agency and value has become a common challenge the international community can no longer ignore. Over millennia of civilizational development, China's fine traditional culture has accumulated valuable insights into humanity's place in the world, the proper bounds of technology, and the underlying dynamics of change. Rooted in historical practice, this Eastern wisdom is increasingly emerging as an important source of values and guidance for the development of a global AI governance system.

Human-centered compass for AI for good

In China's fine traditional culture, the principle that "among Heaven, Earth, and all things, human beings are the most precious" has consistently placed human agency and value at the heart of social development, serving as an important value reference for an AI governance philosophy that is people-centered and oriented toward the good. The Confucian concept of *ren* (benevolence) emphasizes empathy, reciprocity, and authentic human relationships, guiding technological development toward the common good of humanity. AI-generated content can become a socially meaningful product of civilization only through human emotional engagement, value judgment, and intersubjective recognition.

Ming Dynasty (1368–1644) philosopher Wang Yangming's concepts of "attaining conscience through self-reflection" and "unity of knowledge and action" have become important cultural resources for value alignment in the latest generation of AI models. Guiding algorithm design and product development through conscience means enabling intelligent systems to align with humanity's shared moral sensibilities, address real-world challenges such as algorithmic bias and impersonal decision-making, balance technological convenience with human needs, and mitigate the risks of cognitive and emotional imbalance arising from asymmetries in human–machine interaction, thereby making machines more attuned to human values.

The holistic vision of "unity between Heaven and humanity" further broadens the value horizon of a people-centered approach, providing an intellectual foundation for the harmonious coexistence of humans, machines, and nature. Unlike instrumental rationality, which places greater emphasis on the immediate utility of technology, this holistic vision embedded in China's fine traditional culture holds that technological development should not sever the links between humanity and the ecosystem, or between the present and the long-term future.

Dialectical wisdom of 'knowing what to do and what not to do'

As AI becomes capable of performing an ever wider range of tasks, determining whether something should be done is far more fundamental than determining whether it can be done. Past breakthroughs in biotechnology prompted humanity to proactively establish ethical red lines. Today's breakthroughs in AI, however, have more systemic and far-reaching implications. Relying solely on a static list of prohibited areas cannot account for the multitude of unpredictable risks generated by AI's rapid evolution and expanding applications and may also constrain its beneficial development.

Harnessing AI's technological dividends while safeguarding the foundations of human survival and development and preserving human agency and value has become a common challenge the international community can no longer ignore. Photo: CFP

The dialectical wisdom of "knowing what to do and what not to do" in the Confucian tradition offers an approach to AI governance centered on dynamic balance and the careful weighing of multiple considerations. Knowing what to do does not mean acting without restraint. Risk management is a prerequisite for AI development: Every advance in technology should be matched by a corresponding advance in governance. The more powerful the model, the more rigorous the scrutiny of its training data should be; the broader its applications, the earlier their ethical implications should be assessed.

Knowing what not to do, by the same token, does not mean closing the door on progress. Some applications are not permanently off-limits; rather, they should be undertaken once the necessary conditions are in place and adequate safeguards have been established. Such restraint is intended precisely to make action possible over the longer term. When risks become manageable and institutions are able to keep pace, "not to do" can be transformed into "to do."

From governing through shu to governing through dao

Traditional Chinese thought distinguishes between *dao* (the Way) and *shu* (techniques), viewing shu as the tools and means used to address specific problems, and dao as the overarching principles and order that govern the whole. Traditional approaches to technology governance have focused primarily on shu, seeking to maximize the utility of technological tools.

Since the 20th century, the global governance of major technologies has essentially been shaped by competition among actors pursuing their own security and development interests, with equilibrium sought through confrontation and checks and balances. History, however, has repeatedly demonstrated that such equilibrium cannot prevent human conflict. On the contrary, each conflict tends to push existing technological capabilities to their limits.

As AI gains the capacity to "participate" in decision-making and even to "replace" human agency, governance based solely on shu can no longer contain the continuing influence of intelligent algorithms on human cognition or address major risks that threaten the foundations of human civilization. AI governance must therefore shift from governing through shu to governing through dao, grasping the underlying logic of the intelligent age and steering its direction of development.

The impact of AI on human society is diffuse and far-reaching. Continuing to follow a competitive logic in which "whoever achieves breakthroughs first gains the advantage" risks producing a race to the bottom, with safety standards continually lowered and ethical boundaries repeatedly compromised—an outcome detrimental to all.

Yin Jie is Party secretary of Shanxi Normal University.

LI FEIYUE

Digital humanities: Beyond answers, toward new questions

Digital humanities represents an emerging form of the humanities rather than an alternative to traditional humanistic scholarship. New technological approaches should certainly be used to address existing research questions, but more importantly, they can bring into view questions that traditional humanities scholarship has not previously recognized as problems at all. Such questions may have no established empirical precedents to draw upon and may even require a break with conventional assumptions. Digital humanities should therefore be understood as a fallible, incomplete, and updatable mode of inquiry—one better equipped to engage with the future.

Falsifiable assumption

The fundamental cyclical model of traditional humanistic scholarship is the hermeneutic circle—a process in which texts and interpretations continually deepen one another. As interpretations become increasingly coherent, disagreements are gradually resolved. Yet preserving the openness of knowledge requires allowing others to use the same data to falsify existing conclusions. The more sophisticated a model becomes, the more important it is to ask what evidence would cause it to fail; the more internally consistent a research framework is, the more actively counterexamples should be sought. This commitment to falsifiability is a prerequisite for maintaining the vitality of digital humanities. No conclusion derived from data and algorithms should therefore be regarded as an ultimate truth, but rather as a hypothesis with only provisional validity.

By transforming algorithmic black boxes into transparent structures open to questioning and refutation, digital humanities may render some of the questions posed by older paradigms irrelevant altogether. Humanistic scholarship should thus welcome falsification because it signals contact with the genuine boundaries of knowledge.

Incomplete exploration

At the methodological core of digital humanities lie formal systems. When algorithms are used to analyze algorithms themselves, or when formal methods are employed to investigate the limits of formalization, the inquiry enters a Gödelian realm of self-reference.

Kurt Gödel demonstrated that in any sufficiently powerful and consistent formal system, there necessarily exists a proposition that can be neither proved nor disproved. Any attempt to address the future through a single algorithm will eventually encounter the limits of computability. Meaning, emotion, aesthetic experience, and historical context are inherently resistant to complete formalization. The same text may generate radically different interpretations across historical contexts and readerships. Aesthetic judgment cannot be fully translated into computable feature vectors; formalization will inevitably leave something unresolved and encounter cases that cannot be decided algorithmically.

Measured against the full body of historical texts, any database is necessarily an incomplete sample. The very act of indexing a text already reflects the researcher's assumptions and limitations. To acknowledge "incompleteness" is to recognize that humanistic understanding is always a work in progress. Such an acknowledgment rejects the idea that knowledge can be enclosed within a self-sufficient system. Instead, it calls for sustained attention to the boundaries of formalization and to the tension between the computable and the uncomputable, allowing digital humanities to develop as a genuinely open, future-oriented form of inquiry.

Updatable cycle

The basic logic of Bayes' theorem is to begin with a hypothesis, allow the evidence to speak, and revise that hypothesis as new evidence emerges. The initial hypothesis may be wrong, and some degree of error at the outset is unavoidable. What matters is the speed and magnitude of revision: Each new piece of evidence requires the relative weights assigned to different possibilities to be recalibrated. Humanistic data often present the dual challenges of small sample sizes and high dimensionality. Some datasets may contain only a few dozen texts, while each one may involve thousands of feature dimensions.

The "priors" of humanistic scholars often derive from beliefs and experience, scholarly traditions, and historical contexts that are difficult to quantify. Digital humanities seeks to make such tacit knowledge explicit. Not all evidence is equivalent: Evidence that merely confirms existing preferences must be distinguished from evidence that genuinely challenges a hypothesis. At the same time, conclusions should be probabilistic, carry degrees of confidence, and remain open to revision. Once the evidence becomes sufficiently compelling, a fundamental change in belief is not a betrayal but a rational response. Updatability means that scholarly findings are no longer one-time deliverables, but objects subject to continuous revision and maintenance.

By introducing mechanisms for detecting errors, acknowledging limitations, and self-correction, digital humanities incorporates its own epistemological limits into its methodology. Rather than accepting the fixed boundaries implied by "this is how it has always been" or "this is where it ends," digital humanities operates within an as-yet-unnamed territory between computation and interpretation, probability and narrative, and the decidable and the undecidable—one in which every line of code and every annotation remains open to revision.

Li Feiyue is a professor from the School of Humanities at Tsinghua University.