Tomorrow's Affairs
Analysis of today
Assessment of tomorrow

By: Ana Palacio

AI governance - striking the balance between innovation and safety

One way to represent the AI age is with a steep curve and a gradual slope. The curve rises almost vertically: models rapidly grow cheaper, more powerful, more accessible, and more deeply embedded in production, science, education, finance, administration, security, and war.

Beneath it, the slope moves on a far shallower incline. This is the capacity of society to absorb change.

The slope encompasses updates to regulatory regimes, legal structures, and the competencies of public administrations.

It also includes changes to school curricula, business models, workflows, and jobs. And it covers citizens' ability to function in this new landscape—including their capacity to distinguish truth from simulation, persuasion from manipulation, and assistance from dependency.

It is the social metabolism through which a technological shock is integrated into a functioning order.

To be sure, previous technological and industrial revolutions could also be represented by such a curve and slope.

Technologies advanced and spread faster than regulations, business models, education systems, labor markets, and lifestyles could adapt.

Transitions were often painful and never automatic. But the gap between curve and slope was smaller: in speed and scale, the AI-driven transformation appears to be in a league of its own.

Discussion about how to manage this transition usually boils down to the question of how much regulation strikes the "right balance" between innovation and safety.

We do not want to put a bureaucratic lid on progress any more than we want to leave societies to absorb, alone and at speed, the consequences of a technology that is transforming our most basic understanding of work, knowledge, power, accountability, and judgment.

But this question is too narrow. Regulation has a crucial role to play, but it alone does not amount to governance. For that, we must determine the conditions needed to enable societies to live with AI without being overrun by it.

AI is not intangible

The first step is to stop thinking of AI as intangible—some ethereal technology that exists in the "cloud" and delivers whatever the prompt demands.

AI depends on tremendous quantities of capital, cutting-edge chips, cybersecurity systems, supply chains, and talent. It also relies on massive data centers, which emit noise, light, and heat, and require land and permits, as well as huge amounts of electricity and water.

> Difficult decisions about zoning, electricity prices, and environmental protection must be made

The communities hosting the physical foundations of this supposedly weightless technology are now being forced to confront major disruptions and bear high costs, as their lands are bought, their grids are strained, and their water supplies are drained and polluted.

Difficult decisions about zoning, electricity prices, and environmental protection must be made.

The stakes feel existential

But AI governance is not just a community-level challenge. The technology is increasingly central to the infrastructure of national power: military targeting, cyber operations,

intelligence analysis, scientific discovery, industrial automation, surveillance, financial modeling, and political influence.

Whoever controls the most advanced models, chips, compute clusters, and talent holds not merely a commercial advantage, but a strategic one.

This is not lost on the United States and China, which are competing doggedly for AI leadership: each believes that falling behind would leave it militarily, economically, or politically vulnerable.

> Much like with nuclear weapons, the stakes feel existential

Much like with nuclear weapons, the stakes feel existential. This is a recipe for secrecy, mistrust, pre-emption, and unchecked acceleration.

The risks are compounded by the fact that AI is more diffuse than nuclear technology, more commercially embedded, and more broadly applicable.

Moreover, its output is hard to verify, and once capabilities are encoded in software, proliferation is easy.

An 'AI Paris Agreement'

AI governance must account for all of these factors, from the technology's material requirements to its effects on individuals, communities, businesses, labor markets, the environment, national security, and international strategic competition.

To this end, a comprehensive architecture of verification-based trust—broader than national AI regulation and more flexible than a classic treaty—is essential.

The 2015 Paris climate agreement offers a useful model. It created a living framework encompassing voluntary commitments, mutual pressure, regular updates, reliable measurement, and broad participation.

The 2015 Paris climate agreement offers a useful model. It created a living framework encompassing voluntary commitments, mutual pressure, regular updates, reliable measurement, and broad participation

But an analogous AI agreement must also reflect a key lesson of non-proliferation regimes: where political trust is weak, it must be supplemented by testing, limits, and verification.

The resulting framework would combine the flexibility of climate diplomacy with the discipline of nuclear control.

Evolving commitments would be matched by mechanisms for assessing the credibility and impact of those commitments. State action would be accompanied by responsibilities for corporations.

An "AI Paris Agreement" would provide common definitions for frontier models; thresholds for enhanced scrutiny; mandatory incident reporting; shared evaluation standards; protection of model weights; rules for deployment in sensitive sectors; and independent testing of systems that pose risks to cybersecurity, biological safety, critical infrastructure, or military decision-making.

The curve will keep getting steeper

Each of these elements would be regularly reviewed and strengthened over time, by technical bodies that understand the evolving technology and political bodies capable of imposing consequences when commitments are ignored.

Of course, not every application requires the same treatment. A narrow tool used in low-risk settings requires limited oversight.

By contrast, a frontier model capable of affecting cybersecurity, biological design, military operations, or the autonomy of critical systems must be subject to stricter requirements and, in extreme cases, intervention.

> No one would board an aircraft or take a new drug if safety depended solely on the goodwill of manufacturers

Some might argue that this would stifle innovation. But no one would board an aircraft or take a new drug if safety depended solely on the goodwill of manufacturers. In these fields, regulation does not eliminate risk; it makes risk socially bearable. The same is true for AI.

Such a framework cannot be presented as a European project, an American instrument, or an alliance aimed at countering China.

Its legitimacy would depend on inclusion, with countries at the technological frontier, "middle powers," and developing economies all given a voice.

Universities, companies, and civil-society organizations should also be involved.

The curve will keep getting steeper. We cannot flatten it by decree, nor obscure it with speeches.

But we can decide whether it rises ungoverned or within an architecture that can keep up.

Ana Palacio, a former minister of foreign affairs of Spain and former senior vice president and general counsel of the World Bank Group, is a visiting lecturer at Georgetown University.