



By: Gordon LaForge - Robert C. Orr

Could middle powers govern frontier AI development?



The warnings about the safety of frontier AI models continue to grow louder. Writing in Time in early October, the former chief scientist of the United Kingdom's **AI Security Institute** put the probability of us all dying because of smarter-than-human AI systems at 50%.

On the same day, a senior **OpenAI employee** announced that he had left the company because of its “broken” safety culture.

But the United States has shown no interest in regulating the technology, except for a “morally binding” voluntary accord recently signed by tech CEOs and US President Donald Trump.

This raises the question of who could govern frontier AI development. A group of middle powers displayed a willingness to do so at the United Nations General Assembly in September, fresh off the news that OpenAI agents had escaped their testing environment and hacked into Hugging Face and Australia's national health-care database.

Finnish President Alexander Stubb and Norwegian Prime Minister Jonas Gahr Støre launched A **Call for Control of Frontier AI Models**, which has been endorsed by more than 30 political leaders and calls for more coordination and the creation of an international AI agency.

To be sure, such declarations rarely lead to action. But this moment could be different.

Market size

AI risk has taken over the public imagination globally, creating an opening for other countries to shape the terms on which the frontier is developed and deployed.

Doing so, however, requires banding together to convert their collective market, regulatory, and diplomatic weight into strategic leverage.

Middle powers' most powerful lever is market size, as AI frontier labs compete to acquire

enterprise and consumer customers outside of the US.

A sufficiently large coalition of countries—say, India, Japan, South Korea, Brazil, South Africa, and European Union member states—could condition market access on meeting a common set of safety requirements that might include mandatory pre-deployment model testing by a national AI safety institute or third-party auditors.

While few countries can afford to test frontier models on their own, that could be addressed by joint investment in technical capacity

While few countries can afford to test frontier models on their own, that could be addressed by joint investment in technical capacity.

Pooling resources and channeling them to institutions like the UK's AI Security Institute would give each coalition member access to independent testing facilities.

Moreover, shared incident monitoring and reporting could provide early warnings of model failures and generate evidence for future rules.

Safety agenda

Looking ahead, a coalition of middle powers should align on an ambitious safety agenda for the Geneva AI Summit in June 2027.

Loss-of-control incidents are likely to have worsened by then, and recursive self-improvement may be fully realized, putting the international community under immense pressure to rein in the industry.

Middle powers should be ready to demand stronger risk-management practices from tech companies

Middle powers should be ready to demand stronger risk-management practices from tech companies and present a shared set of **AI red lines**. Even if those demands are not legally enforceable, they could begin to shape global norms and national policies.

It will be difficult for middle powers to find common ground on these issues, given their diverse development levels and attitudes toward AI.

Some will invariably fear losing access to frontier models or compute—America's dominance in AI gives it substantial leverage over these countries. But that is all the more reason for them to align their positions and engage with the great powers as a united front.

Coalitions of national and international policymakers

This is not without precedent. Countries have come together to address nuclear safety, pandemic preparedness, counterterrorism, and other consequential issues.

For example, the Paris climate agreement would not have been possible without middle powers and the most climate-vulnerable countries uniting to force the US and China to the negotiating table.



The Paris climate agreement would not have been possible without middle powers

Moreover, coalitions of national and international policymakers, governors, mayors, CEOs, researchers, and civil-society groups created demand for climate solutions—including renewables, now the

fastest-growing **energy source** and expected to overtake coal this year in global electricity generation. Middle powers were instrumental in triggering this historic shift.

Of course, ensuring AI safety will require US regulation and an agreement between the US and China on loss-of-control, cybersecurity, and military uses of the technology.

But AI development and deployment will be safer and more beneficial if a wider range of countries can help shape it.

Middle powers have already influenced AI safety in important ways: the UK's 2023 AI Safety Summit spurred Britain and the United States to launch government AI safety institutes, and South Korea's follow-up summit in 2024 produced an international network of them.

Now they must build on these efforts to put all countries, including the superpowers, on the path toward safer AI.

Gordon LaForge is Co-Director of the Power, People, and Planet Initiative at New America and Senior Fellow for Global AI Policy at the Starling Institute.

Robert C. Orr, a former United Nations assistant secretary-general for strategic planning, is Special Adviser to the UN Secretary-General for Climate Change and Director of the Global Policy Institute at the University of Maryland.