The Compounding Error

Frontier AI is accumulating capability and error in equal measure. The next strategic advantage will belong to those who can measure, manage and repay that error debt.

Earlier this year in Tallinn, analysts at Estonia's Foreign Intelligence Service ran an unusual interrogation. They tested DeepSeek, China's flagship open-weights model, on questions of Baltic security, and their conclusion deserved far wider attention: the system systematically concealed critical facts and inserted Beijing's political orthodoxies into its answers.

The obvious story was Chinese censorship. The more dangerous story was the technological climate into which that censorship landed.

Artificial intelligence has reached an awkward threshold. It is capable enough that governments, enterprises and citizens are delegating real authority to it, and unreliable enough that its outputs cannot be presumed true. The gap creates a vulnerability democracies have barely begun to price. An adversary no longer needs an expensive propaganda apparatus to convince a foreign public of a falsehood; it needs only to seed engineered distortions into an information ecosystem already churning out convincing mistakes of its own.

The next geopolitical contest over artificial intelligence will not be fought over raw capability. It will be fought over verification.

Today, the AI race is measured in compute clusters, parameter counts and benchmark leaderboards, underpinned by a quiet assumption that reliability will scale alongside capability. It has not. Intelligence and reliability are distinct engineering challenges, and a system that sounds brilliant is not thereby trustworthy. OpenAI's own testing of its o3 reasoning model recorded hallucination on 33 per cent of attempted answers about real people, double the rate recorded for its predecessor. Its smaller sibling, o4-mini, fabricated answers nearly half the time, at 48 per cent. Stanford's 2026 AI Index reveals the same jagged frontier: models capable of winning gold at the International Mathematical Olympiad yet unable to reliably read an analogue clock, and autonomous agents that fail complex computer workflows roughly once in every three attempts.

A single failure in a drafted email is an editorial annoyance. But modern systems are increasingly autonomous and sequential: an agent retrieves a file, writes a script, executes code, queries a database, authorises an action and feeds the result into its next step. The mathematics is unforgiving. A twenty-step workflow in which every step succeeds 98 per cent of the time completes without error just 67 per cent of the time. Capability scales linearly. Error compounds exponentially.

The compounding also operates across time. As machine-generated text floods the public web, it is ingested as training data by the next generation of models. Researchers evaluating 800,000 synthetic medical data points showed the consequences when that feedback loop operates without human oversight: pathological diversity erodes, false reassurance rates triple to 40 per cent, rare but lethal conditions quietly vanish, and, within two generations, blinded physicians rate the records as clinically useless. This is error debt; a hidden liability incurred whenever an unverified machine output enters a legal brief, medical file, intelligence assessment or training corpus. Like financial debt, it accumulates silently but must eventually be repaid through costly forensic verification. Left unchecked, it erodes confidence in the record itself and ultimately makes the past impossible to reconstruct.

> Error debt accrues silently: every unvetted machine output that enters a legal brief, medical file, intelligence assessment or training corpus is a liability someone must eventually repay.

For hostile intelligence services, error debt is not a flaw to be exploited opportunistically; it is an attack surface by design. Information warfare once demanded industrial effort: inventing narratives, constructing distribution networks and overpowering independent journalism. Generative AI has transformed those economics. Commercial models already produce convincing stochastic errors, fabrications generated by their probabilistic architecture, so an adversary need only seed the information environment with carefully engineered

falsehoods. Corrupting a rival's decision-making has never been cheaper. More importantly, the prevalence of ordinary machine error delivers something classical propaganda never offered: plausible deniability. When a planted falsehood appears in an allied intelligence assessment, it is virtually impossible to distinguish from a routine AI hallucination.

Beijing has understood the opportunity. Swedish and European government researchers have documented content controls across DeepSeek, Alibaba's Qwen and Moonshot's Kimi that steer users towards party-approved narratives far beyond China's borders. These open-weight models, deliberately engineered to be inexpensive and widely deployable, are spreading through governments and enterprises across Southeast Asia, Africa and Latin America. Yet Western systems are not insulated from the same dynamics. A peer reviewed study in Nature this year traced Chinese state media content deep into standard pre-training datasets and found that leading commercial models generated significantly more favourable depictions of Chinese institutions when queried in Mandarin than in English. The pattern extended across 37 countries: models consistently portrayed governments more positively where media systems were more tightly controlled, particularly in those governments' own languages. This is information warfare moving upstream. The contest is no longer over the information citizens consume, but over the computational infrastructure through which organisations, militaries and states interpret reality.

In that environment, the democratic world's decisive advantage is not informational purity but institutional correctability. Open science, adversarial journalism, independent courts, whistleblowers and competing laboratories constitute an error-correction architecture unmatched by authoritarian systems. When courts in the United States and Australia sanctioned lawyers for submitting machine-fabricated citations, the mechanism worked exactly as intended: error surfaced, accountability followed and confidence in the system was reinforced. Authoritarian regimes cannot reproduce this dynamic because political legitimacy takes precedence over factual correction. One system can reduce its error debt. The other is condemned to refinance it indefinitely.

Quantum computing offers the telling blueprint. For thirty years the field treated error as its foundational bottleneck, and when Google's Willow processor demonstrated in Nature that logical error rates could fall as quantum systems scale, it achieved that milestone before claiming practical usefulness. Critics dismissed the discipline as slow-moving. Yet that caution produced something remarkable: the first class of computer whose reliability strengthens as it grows. Artificial intelligence pursued the opposite strategy. The world scaled deployment first and only afterwards began grappling with reliability, doing so in live systems that increasingly shape economic, social and geopolitical decisions.

Reversing that logic requires treating verification as critical national infrastructure. The policy agenda is neither abstract nor distant. Governments should reward

demonstrated reliability rather than benchmark performance, requiring auditable data provenance, continuous red-teaming against model poisoning and clear limits on autonomous self-generation. In intelligence, energy and defence, execution should be separated from verification, with outputs validated through independent architectures rather than accepted on the basis of a model's own confidence score. The growing alliance of AI safety institutes across London, Washington, Tokyo, Seoul, Singapore and Canberra should establish common reliability standards just as the Basel Committee established capital standards for banking, ensuring that error debt, like financial debt, is measured, disclosed and managed before it becomes systemic. Above all, human oversight must be understood correctly. Keeping people in the loop is not a temporary safeguard while the technology matures. It is the principal mechanism through which error debt is paid down before compound interest takes hold.

The public has already sensed what policymakers have been slow to formalise. A 47 country study by the University of Melbourne and KPMG found 66 per cent of people now use AI regularly while just 46 per cent are willing to trust it, and 70 per cent believe regulation is needed. Citizens understand instinctively that delegating consequential decisions to machines, without the infrastructure to audit them, is systemic fragility.

The early phase of the AI race rewarded whoever could process information at the greatest velocity. The next phase will belong to those who can prove what is true. In an automated world, epistemic sovereignty - the institutional capacity to verify reality - is the ultimate source of power.