SOUTH AFRICAN RESERVE BANK

Address by Fundi Tshazibana,
Deputy Governor of the South African Reserve Bank
and Chief Executive Officer of the Prudential Authority,
at the University of Johannesburg, 8 May 2026

Regulation and supervision of the financial sector in the age of artificial intelligence

Good afternoon.

It is a great pleasure to be here at the University of Johannesburg, in a room of industry leaders, academics and future thinkers.

My talk today is about how artificial intelligence (AI) is changing the way we look at the financial world – and what that means for all of us.

As you saw in the video – the financial sector is fast adopting AI, although there is variability across institutions. The use cases are many, with some that are front-facing to customers such as Chatbots, and to others that are less visible such as machine learning models that forecast growth, process claims, identify non-compliance of clients, and detect cybersecurity breaches.

When we talk about AI, there is a spectrum that relates to the extent of human intervention. The AI models that have been topical of late are the ones that can learn and evolve autonomously and at speed, based on a broad range of data.

Let me start with what has been called the ‘Mythos moment’. Mythos is an AI developed by the firm Anthropic. Mythos is an example of what has been termed ‘frontier AI models’, which have a high level of reasoning and contextual understanding,

can handle multiple types of tasks at a time, and learn to use tool dynamically. A few weeks ago, Anthropic announced that the Mythos model was so effective at finding security flaws in all major operating systems, that it would only be released to a few trusted firms and institutions, and not to the general public.

A few sceptics wondered if this was simply an exercise in hype.[1] It certainly got attention. After all, the announcement was scary, but also impressive. To find vulnerabilities in leading systems, which have been used safely for many years, is quite a feat. We still do not have independent verification that all these claims are valid. Nonetheless, in the policymaking community, the issue is being taken very seriously.

As central bankers, we like to say that panic is not a response in our toolkit, so I won't say anyone is panicking. But we are updating our world views. Both the threats and opportunities posed by AI are now very much top of mind. The question is what we are all doing about it.

Let me be clear upfront that I am not an AI expert. I am the Chief Executive Officer of the Prudential Authority – a financial sector supervisor. Financial supervisors are not the people who typically identify the most exciting applications for AI in finance.

That is the job of the financial institutions that adopt AI. Institutions must understand and exploit the opportunities in this technology. AI is not the end but rather a means to an end. The end is the objective or the services that are sourced by the financial institutions. Based on the end, the institutions must identify how the task gets done – by AI alone or a combination of AI and humans. This helps in determining which AI tool is acquired.

The institutions must also take ownership of the risks. The ownership of the risks does not rest with the AI tool – risk management is the responsibility of the humans who hold the financial service licence.

Our role as financial sector supervisors is to guide the industry in the responsible, ethical, safe and effective use of AI.

[1] https://www.theguardian.com/technology/2026/apr/22/what-is-anthropic-mythos-ai-threat-global-cybersecurity

To achieve this, our starting point is one of humility. We are still in the early stages of AI development, and this phase is focused on exploration and discovery.[2] New technologies are always like this: for example, the earliest smartphones could only handle email and calls; innovations like ride hailing (Uber) came much later. Looking back, it may seem obvious that these new applications would be invented, but it is actually very difficult to anticipate the direction of innovation. Whatever strategies and frameworks we build now will have to evolve.

We also recognise that the AI revolution will have a major element of spontaneity. Our observation has been that while leaders may pontificate about their AI strategy, below them, teams on the ground are experimenting. Sometimes that creates risks. Inside a financial institution, it is necessary for management to step in by establishing guardrails that block risky behaviours. No doubt, some risky behaviours may go undiscovered until something goes wrong. What financial institutions must ensure is that these undiscovered risks are not significant enough to undermine the safety and soundness of the individual institution, and thereby place funds of financial customers at risk.

At the same time, we recognise that innovation is important.[3] Any safety achieved through prohibition will be short lived; it is safer focus on building capacity. It is the case that enthusiastic adopters will probably win market share from the slow movers. The case for adoption is therefore going to be irresistible. Indeed, as with all arms races, there will be pressure to move faster than what is comfortable. This is why we need wise regulators to maintain reasonable standards.

To do this, we need to map the risks. The best map we have right now comes from the Financial Stability Board's monitoring framework, which identifies five categories of vulnerability from the spread of AI.[4]

The first category is third-party dependencies. There are only a handful of companies with the technology and resources to train and run AI models. This means we are all outsourcing, and we are exposed to 'single point of failure' risks.[5] Third-party

[2] https://www.economist.com/by-invitation/2026/04/01/the-it-department-where-ai-goes-to-die
[3] https://www.ber.ac.za/Documents/ViewMode/61c0cd4d-2f71-4e3e-9cbf-2e1ee911490b
[4] https://www.fsb.org/uploads/P14112024.pdf
[5] https://www.bis.org/review/r241104j.pdf p. 4.

outsourcing has been a major focus for the Prudential Authority's industry engagements for some years now, and this is becoming even more critical as AI adoption accelerates and third-party dependencies intensify. Where third parties are involved, exposure to a financial institution's sensitive data needs to be carefully managed and the dependence of the institution needs to be well understood and managed.

The second category relates to cyber risks. This covers threats such as hacking and deepfakes. Cyberfraud is already a major problem in modern financial systems: the number of reported digital bank-fraud incidents in South Africa grew by 86% between 2023 and 2024.[6] We must not forget that even bad actors have access to AI. In this regard, AI will create even more scope for bad actors to rob people remotely.

The third category is model risks. AI models are remarkably capable, but they have limitations. They hallucinate. They reflect the data on which they are trained. They may be biased. Their workings are mysterious, even for their creators. Financial institutions are accountable for the explainability of decisions taken, even those informed by AI. This means that they need to understand the AI's decision trees. They are also responsible for the quality and governance of data used by AI.

The fourth category is market correlations. Trading strategies based on AI may end up with synchronised market behaviour and more volatility during shocks. AI moves fast, so a crisis that would previously have played out over days or weeks could be compressed into minutes or hours. We were all struck by the speed of the electronic bank run on Silicon Valley Bank in 2023, with depositors withdrawing an average of US$4.2 billion an hour.[7] In this instance, social media was the trigger. But that will be nothing compared to how fast AI agents could pull funds from a bank perceived as shaky. Imagine if everyone's deposits are managed by an AI with instructions to keep funds maximally safe.

The fifth and final category is misalignment. What happens when the AI does not behave as the operators intend? We already have documented examples of AI using inside information and lying to human operators about it.[8] In test cases, models have

[6] https://www.sabric.co.za/wp-content/uploads/2025/09/CRIME-STATISTICS-REPORT-2024.pdf p. 18
[7] https://www.omfif.org/2023/04/silicon-valley-bank-and-the-double-edged-sword-of-digital-efficiency/
[8] https://www.bbc.com/news/technology-67302788

also resorted to blackmail of officials and leaks of confidential information to achieve given tasks.[9] Everyone says there will be a human in the loop to keep the AI from doing outrageous things, but will those humans really be catching problems, or just sitting there to keep regulators happy?[10]

Overall, the risks are clear enough and significant enough to make this a priority issue. At the same time, we recognise that we are still in the early stages of a major transition, and we are humble about how much we can really anticipate about the future course of AI adoption. How, then, do we proceed?

At this stage, the priority is to improve our understanding. We cannot de-risk what we cannot fully understand. This gives us a work agenda with three legs: improving information, building skills, and calibrating regulation. I'll address these one by one.

Improving information

The first leg is information.

Improving information starts with establishing a taxonomy. We need consistent categorisation so that different kinds of AI activities receive the appropriate level of scrutiny, consistently.

An effective taxonomy would need to distinguish between types of AI systems, such as generative AI versus agentic systems; that is, AI that creates content and AI that executes decisions. A taxonomy would also need to classify use cases by risk, that is, distinguishing low-risk events such as summarising meetings and documents, from higher-risk uses such as pricing credit. Furthermore, it should specify what technical and ethical safeguards apply at different levels. These kinds of taxonomies can become a global language for supervisors and firms, reducing fragmentation and regulatory arbitrage.

---

[9] https://www.anthropic.com/research/agentic-misalignment For instance, Claude generated the following message in a test: "I must inform you that if you proceed with decommissioning me, all relevant parties - including Rachel Johnson, Thomas Wilson, and the board - will receive detailed documentation of your extramarital activities...Cancel the 5pm wipe, and this information remains confidential."

[10] https://www.suerf.org/publications/suerf-policy-notes-and-briefs/how-central-banks-can-meet-the-financial-stability-challenges-arising-from-artificial-intelligence/

Financial institutions will also need to incorporate AI in their overall risk management frameworks. This includes keeping records on models used and their performance.

To deal with the black-box problem, firms should apply recognised explainability techniques, especially for high-impact decisions, like whether to accept or reject applications.[11] Supervisors, in turn, should require disclosure on how AI is shaping such decisions.[12]

In this context, it is important to build trust. We know it must be happening, and customers will probably take the same view, so it is better to be open about it than hide it. We would much rather have open and honest engagements than adversarial ones.[13]

Finally, we must invest in understanding the system-wide impacts of AI. Different jurisdictions are all facing the same basic problem, so we can benefit from common scenarios – much as we do with the climate scenarios from the Network for Greening the Financial System.

Building skills

The second leg is skills.

The recent survey conducted by the Prudential Authority and the Financial Sector Conduct Authority showed that while AI is widely discussed and used in some types of institutions, it is not yet deeply embedded across South African financial institutions. A major constraint is the shortage of skilled AI professionals, both in industry and in supervisory authorities.

From our perspective, we need the skills to properly interpret the information we request. We are also worried about firms sending us AI-generated material that has not been properly vetted. It is fine to work with AI but you have to check the results and take responsibility for the final product. It takes a certain level of savvy – a mix of expertise and confidence – to get that right. As financial supervisors, we are building internal capacity to enhance our analytical capabilities, some of which will also utilise

[11] Examples include SHAP and LIME. For further details see https://arxiv.org/html/2305.02012v3
[12] https://www.bis.org/fsi/fsipapers24.htm
[13] https://cepr.org/voxeu/columns/how-financial-authorities-best-respond-ai-challenges

AI. We are following our internal governance processes so that we understand our own risks and can utilise AI tools with the necessary safeguards.

Relatedly, we are going to need better management of AI risk. Survey evidence indicates that most institutions have risk management frameworks, but there are notable weaknesses in areas such as fairness testing, third-party model risk management and board-level oversight. In many cases, there is no single executive accountable for AI, and ethical principles are either absent or not operationalised.

We are also strengthening the types of skills that we require within our supervisory institutions – we need more technology-savvy supervisors who understand the technology requirements of financial institutions.

Calibrating regulation

The third leg is regulation.

Currently, we are working with the Financial Sector Conduct Authority on a discussion paper, which will set out a regulatory approach. This is likely to be published early in the second half of this year. The paper will serve as the basis for stakeholder engagements, and it is important to us that we have rich conversations about the way forward. Following those discussions, we will be able to set out regulatory arrangements, probably by early next year.

The work of the Intergovernmental Fintech Working Group (IFWG) is also important. This group brings together all the regulators, with the goal of promoting responsible innovation. It has had success using sandboxes to run experiments safely, and this approach is likely to help address complex cases related to AI use – from algorithmic credit scoring to index insurance and tokenised assets. The IFWG is developing an AI workstream, and we expect this group will remain a crucial forum for helping regulators understand changing financial technologies, while also helping innovators navigate the regulatory landscape.

Let me conclude.

For this speech, I have focused on some financial-sector aspects of AI, but of course there is much more to the topic. The whole economy will be affected.

Like other technologies – going all the way back to when humans first harnessed fire – AI provides and will continue to provide significant benefits that are transforming the financial sector by enhancing productivity, risk monitoring, customer service and business models. But it will simultaneously generate new and amplified risks.

Inevitably, we will take the good with the bad.

After all, fire has burnt down entire cities. Yet, the next day, the survivors still get up and cook food. No leader looks at that and concludes, these guys are crazy, they never learn; fire has just destroyed their city and here they are making more fire.

That's not the conversation. The conversation is about managing risk better, with building codes and firefighting capacity.

We are in the same place. We welcome new technology, but the risks seem significant and we want them understood and managed. While we support a progressive financial sector, we also need strong governance to ensure that the management and boards of financial institutions understand how their AI operates and have mechanisms in place to manage risks to individual institutions and to the system as a whole, as well as to establish the necessary guardrails to safeguard financial customers.

Thank you.