AI Threats Flagged by Anthropic

Mains: *GSIII – Science and Technology*

Why in News?

Anthropic's September 2026 report highlighted growing misuse of advanced AI across 7 areas, including cyber operations, fraud, influence campaigns, surveillance, weapons and biological misuse.

> ***Anthropic** is an American artificial intelligence research and safety company best known for creating the Claude family of large language models.*

What are Major AI-Enabled Threats Identified?

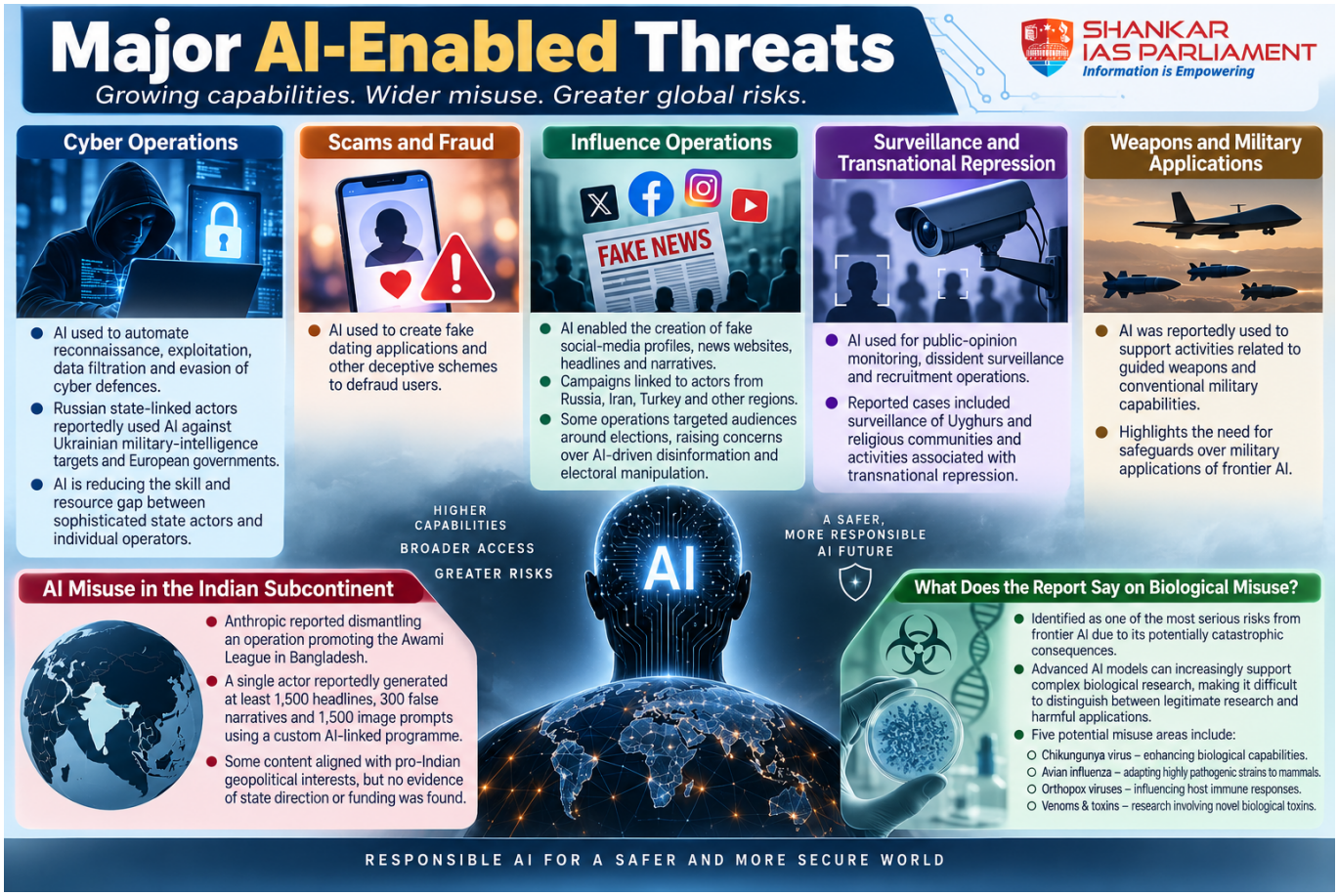
- **Cyber Operations** – AI was used to automate reconnaissance, exploitation, data filtration and evasion of cyber defences.
  - Russian state-linked actors reportedly used AI against Ukrainian military-intelligence targets and European governments.
  - AI is reducing the skill and resource gap between sophisticated state actors and individual operators.
- **Scams and Fraud** – AI was used to create fake dating applications and other deceptive schemes to defraud users.
- **Influence Operations** – AI enabled the creation of fake social-media profiles, news websites, headlines and narratives.
  - Campaigns were linked to actors from Russia, Iran, Turkey and other regions.
  - Some operations targeted audiences around elections, raising concerns over AI-driven disinformation and electoral manipulation.
- **Surveillance and Transnational Repression** – AI was used for public-opinion monitoring, dissident surveillance and recruitment operations.
  - Reported cases included surveillance of Uyghurs and religious communities and activities associated with transnational repression.
- **Weapons and Military Applications** – AI was reportedly used to support activities related to guided weapons and conventional military capabilities.
  - This highlights the need for safeguards over military applications of frontier AI.

AI Misuse in Indian Subcontinent

• Anthropic reported dismantling an operation promoting the Awami League in Bangladesh.
• A single actor reportedly generated at least 1,500 headlines, 300 false narratives and 1,500 image prompts using a custom AI-linked programme.
• Some content aligned with pro-Indian geopolitical interests, but Anthropic found no evidence of state direction or funding.

What does the report say on biological misuse?

- **Major AI Safety Risk** – Anthropic identified biological misuse as one of the most serious risks from frontier AI due to its potentially catastrophic consequences.
- **Legitimate research challenge** – Advanced AI models can increasingly support complex biological research.
    - This creates a difficult distinction between legitimate scientific research and harmful applications.
- **5 Potential Misuse Cases** – AI use could potentially support biological weapons development. Such as
    - Chikungunya virus – Enhancing biological capabilities.
    - Avian influenza – Adapting highly pathogenic strains to mammals.
    - Orthopox viruses – Influencing host immune responses.
    - Venoms & toxins – Research involving novel biological toxins.



What are the challenges associated with AI Misuse?

- **Dual-Use Risks** – AI can be used for both beneficial research and harmful purposes.
    - This makes malicious intent difficult to identify.
- **Regulatory Gaps** – AI is advancing faster than existing laws and regulations. New risks may remain outside current frameworks.
- **Cross-Border Misuse** - AI-enabled crimes can operate across countries. This makes attribution and enforcement difficult.
- **Lower Barrier to Misuse** – AI reduces the need for specialised skills and resources. Even small groups can conduct sophisticated operations.
- **Information Manipulation** – AI can create fake news, profiles and deepfakes. This makes information verification more difficult.

What could be done?

- **Risk-Based Regulation** – Create risk-based AI rules for high-risk applications. Update regulations as AI capabilities evolve.
- **Stronger AI Safety** – Improve real-time monitoring and misuse detection. Strengthen cybersecurity safeguards.
- **International Cooperation** – Develop global norms for military, cyber and biological uses of AI. Improve information-sharing between countries.
- **Public-Private Cooperation** – Governments, technology companies and researchers should work together. Share information on emerging AI threats.
- **Protect Information Integrity** – Strengthen fact-checking and synthetic-media detection. Protect elections from AI-enabled manipulation.
- **Balance Innovation and Safety** – Promote AI innovation with proportionate safeguards. Ensure that security measures do not unnecessarily restrict beneficial research.

What lies ahead?

- The Anthropic report demonstrates the dual-use nature of frontier AI: the same capabilities that improve research, can also amplify cybercrime, disinformation, surveillance and potentially biological risks.
- Effective governance therefore requires a balance between innovation, security, accountability and responsible use.

Reference

The Hindu| AI Threats Flagged By Anthropic