AI Threats – Emerging Risks and the Need for Global Guardrails

Mains: *GS III – Cybersecurity*

Why in News?

A recent report by Anthropic, the US-based AI company highlights how AI is being used as an active instrument in cyber operations, fraud, propaganda, surveillance and potentially dangerous biological research.

What is the Background?

- **Benefits of AI** – Artificial Intelligence (AI) is increasingly transforming governance, economic activity, scientific research and national security.
- **Concerns of Misuses** – However, the rapid improvement in frontier AI models has also lowered the barriers to sophisticated misuse.
- **Report by Anthropic** – The report, covering malicious activities detected and disrupted between December 2025 and August 2026, identifies seven major areas of concern:
    - **Scams and fraud, cyber operations, illicit distillation, influence operations, surveillance, conventional weapons and biological misuse**.
- Its significance lies in demonstrating that AI-related threats are no longer purely hypothetical; they are increasingly becoming operational.
- **From AI Assistance to AI-Enabled Operations –** One of the most important findings is the transition from simple human-AI interaction to **AI-orchestrated operations**. According to Anthropic, several cases involved multi-agent frameworks capable of reconnaissance, data processing, information filtering and exploitation.
- Humans continued to determine targets and review important outputs, but AI performed substantial portions of the operational workload.
- This represents a qualitative shift in the nature of technological risk.
- Earlier, sophisticated cyber operations generally required teams possessing specialised technical knowledge.
- Increasingly capable AI models can compress this expertise into automated workflows.
- Consequently, the distinction between a well-resourced state actor and an individual or small group of operators is becoming less pronounced.
- This democratisation of advanced capabilities can generate enormous benefits, but it can also **democratise malicious capacity**.

What are the concerns raised by the report?

- **Cyber Operations and Espionage –** Cybersecurity emerged as one of the most immediate areas of concern.
- Anthropic reported a case involving activity consistent with **Russian state-linked espionage**, in which AI was used to automate parts of operations targeting Ukrainian military-intelligence interests and European governments.
- AI can potentially accelerate reconnaissance of targets, identification of vulnerabilities, generation and modification of malicious code, evasion of cyber defences, processing of stolen information and coordination of multiple stages of an attack.
- The implication is significant for national security. Cyber conflict may become faster, cheaper and more scalable, making conventional distinctions between state and non-state threats increasingly difficult to maintain.
- **Influence Operations and Disinformation –** Another major threat is the industrialisation of **information warfare**.
- Anthropic identified operations in which AI was allegedly used to create social-media profiles, generate deceptive narratives and develop entire news websites.
- Such campaigns reportedly originated from or were associated with actors in countries including Russia, Iran and Türkiye, as well as actors across regions such as South Asia, Africa and Europe.
- The proximity of some campaigns to elections is particularly concerning.
- AI-generated content can make political manipulation cheaper, faster, multilingual, highly personalized and difficult to distinguish from authentic content.
- The danger is not limited to fake news. AI can create an entire ecosystem of apparently independent accounts, websites and narratives, thereby manufacturing the illusion of public consensus.
- This threatens **electoral integrity, social cohesion and public trust in democratic institutions**.
- **AI Misuse in the Indian Subcontinent –** The report also highlights the relevance of AI-enabled influence operations to South Asia.
- Anthropic said it dismantled an operation promoting Bangladesh's Awami League. A single actor reportedly generated at least **1,500 headlines, 300 false narratives and 1,500 image prompts** using custom software connected to a large language model.
- Some of the material reportedly aligned with pro-Indian geopolitical interests, although Anthropic found **no evidence of Indian state direction or funding**.
- For India and its neighbourhood, the episode highlights the emerging challenge of **AI-enabled political manipulation**. With elections becoming increasingly digital, synthetic narratives can potentially influence voters, deepen communal or political polarisation and complicate the work of election authorities.
- **Surveillance and Transnational Repression –** Anthropic also reported several operations involving surveillance and monitoring.
- These included public-opinion monitoring, dissident surveillance, recruitment targeting Uyghurs in Syria, intelligence activities concerning religious communities, and campaigns associated with stability-maintenance surveillance and transnational repression.

- AI can make surveillance significantly more powerful by enabling the rapid analysis of enormous quantities of social-media data, images and videos, communications, behavioural patterns and publicly available information.
- The central concern is therefore not technology alone but **technology combined with state power**.
- Without legal safeguards, AI-enabled surveillance could undermine privacy, freedom of expression, religious freedom and political dissent.
- **Biological Misuse –** Perhaps the most consequential aspect of the report concerns **biological misuse**.
- AI has enormous positive potential in biology, including drug discovery, protein research, disease modelling and public-health innovation.
- However, the same capabilities can have dual-use applications.
- Anthropic noted that earlier AI models were evidently too limited to provide meaningful assistance for dangerous biological research.
- With current frontier models capable of participating in complex scientific workflows, that distinction has become less clear.
- The report described five cases involving requests that could potentially support dangerous biological research.
- One case involved a request to help prepare a grant application concerning genetic modification of an organism to enhance biological capabilities relating to chikungunya virus, particularly its transmissibility and immune-evasion properties.
- Another involved research outside the US concerning highly pathogenic avian influenza and its adaptation to mammals and mechanisms of severe disease.
- Other cases concerned research involving orthopox viruses and investigations into novel non-transmissible venoms and toxins.
- Importantly, Anthropic itself cautioned against interpreting these cases as proof that biological attacks are imminent.
- It stated that some interventions were undertaken out of an **abundance of caution** because the evidence of actual dangerous biological use was ambiguous.
- Nevertheless, the cases demonstrate the growing importance of **AI-biosecurity**.

Why AI-Biosecurity Requires Special Attention?

- Biological risks differ from conventional cyber risks because the consequences of successful misuse can extend beyond digital systems to human health and the environment.
- The fundamental challenge is the **dual-use dilemma**: the same scientific knowledge can support both legitimate research and harmful applications.
- Therefore, simply banning AI assistance for broad categories of biological research may impede legitimate scientific progress.
- At the same time, unrestricted access to highly capable models could lower barriers to dangerous experimentation.
- This calls for sophisticated safeguards based on risk, intent, capability and context rather than crude restrictions.

What are the Implications for India?

- India needs stronger **AI-enabled cybersecurity capabilities** across critical infrastructure, defence, banking and public institutions.
- Election authorities and technology platforms need mechanisms to detect coordinated synthetic influence campaigns and deepfakes while protecting legitimate political speech.
- India must develop a robust **AI-biosecurity framework**, particularly as its biotechnology and pharmaceutical sectors expand.
- AI governance must balance innovation with accountability. Excessive regulation can suppress beneficial applications, while weak safeguards can create systemic risks.
- India should strengthen international cooperation because AI-enabled threats are inherently **transnational**. Cyberattacks, disinformation campaigns and biological risks cannot be effectively managed by individual countries acting in isolation.

What could be done?

- **Risk-based AI regulation** – High-risk applications should face stronger testing, monitoring and accountability requirements.
- **AI safety evaluations** – Frontier models should undergo rigorous red-team testing before and after deployment.
- **Human oversight** – Critical decisions involving weapons, biological research, elections and surveillance should not be fully automated.
- **Content provenance** – Digital systems should develop reliable mechanisms to identify AI-generated or manipulated content.
- **AI-biosecurity safeguards** – Models should detect and restrict assistance that could materially facilitate dangerous biological activities.
- **International cooperation** – Countries should work towards common standards on AI safety, cyber operations, autonomous weapons and biological misuse.
- **Responsible innovation** – Regulation should preserve legitimate scientific and technological applications while limiting high-consequence misuse.

What lies ahead?

- For India, the appropriate response is neither technological alarmism nor unrestricted optimism.
- The objective should be to build an ecosystem where **innovation, national security, democratic resilience, privacy and human welfare coexist**.
- As AI systems become more autonomous and capable, the principle of *"human in the loop"* must evolve into **"human responsibility in command."**
- The future of AI governance will ultimately depend not merely on how intelligent machines become, but on how effectively societies govern the power they acquire.

Reference

The Hindu| Anthropic Report on AI

SHANKAR
IAS PARLIAMENT
Information is Empowering